ICIS 2022 Proceedings

Online Reviews and Recommendations

Dec 12th, 12:00 AM

Can 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

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

Completed Research Paper

Yulia Vorotyntseva
Chaifetz School of Business
Saint Louis University
St. Louis, MO
yulia.vorotyntseva@slu.edu

Aleksi Aaltonen
Fox School of Business
Temple University
Philadelphia, PA
aleksi@temple.edu

Subodha Kumar
Fox School of Business
Temple University
Philadelphia, PA
subodha@temple.edu

Paul A. Pavlou
C. T. Bauer College of Business
University of Houston
Houston, TX
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 offers 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 average ratings across platforms influence platform choice and discuss implications to platform owners, regulators, and consumers.

Keywords: Experiment, food ordering, online ratings, platform competition, ratings inflation, spillover effect

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. Most Americans say that they use ratings to inform their purchases (Turner and Rainie 2020), while another recent study found that only 48% of consumers would consider using a local business with a lower than four-star rating (Murphy 2020). Average ratings and qualitative product reviews associated with them allow consumers to compare products in a way that is not possible in brick-and-mortar setting, or using mail order catalogs, which is a major advantage for online retail platforms that have become ever more popular during the global pandemic (Thomas 2021a,b). At the same time, average ratings are known to be vulnerable to various biases including attempts by platform vendors to artificially boost the numbers. This can contribute to ratings inflation threatening to render ratings systems less informative to consumers (Filippas et al. 2020, Kokkodis 2021, Zervas et al. 2020).

A retail platform owner should ideally work against biases including ratings inflation to make sure ratings provide accurate information about the quality of vendors and their items to consumers. However, upward biased average ratings and fake reviews can drive sales which means that it is not always clear what the platform owner will do, especially if there are competing platforms vying for the same consumers (Ananthakrishnan et al. 2020, Aziz et al. 2020, Rietveld and Schilling 2020, Wang et al. 2020). Should the platform owner fight the biases or let the ratings become inflated to avoid being hurt in an increasingly competitive environment? This is an important dilemma that affects how well retail platforms can match supply and demand in different settings. Food ordering, ride sharing, daily deals, and lodging are just a few examples of platform-based markets in which consumers can choose between alternative platforms that have many of the same vendors and items present. Recent studies suggest that if both consumers and vendors multihome, that is, use several platforms simultaneously, then the opportunities of the platform owner to create competitive bottlenecks and capture value from platform interactions can become significantly more precarious than if the platform acts as the sole gatekeeper between the parties (Bakos and Halaburda 2020, Kim et al. 2017, Rietveld and Schilling 2020). Importantly, a platform owner may have to consider the threat that consumers take their transactions to a competing platform.

Intuitively, a retail platform owner may not want to let their vendors look better on competing platforms. A few studies suggest that average ratings may influence competition between platforms (Kathuria and Lai 2018, Zervas et al. 2020) and, for instance, Apple and Google have made it reciprocally easier for developers to implement designs to boost app ratings on iOS and Android platforms (McGee 2020). Yet, we do not know if there is a real threat from hosting lower average ratings than the competitor or whether such a threat is merely an unfounded perception. This is due to the almost complete lack of studies trying to identify mechanisms by which average ratings could impact platform competition. What we know is that vendors tend to develop legitimate, semi-legitimate and outright fraudulent tactics to boost their ratings, to which platform owners may (or may not) respond with various countermeasures (Ananthakrishnan et al. 2020, de Langhe et al. 2016). The important unanswered question is whether there is an actual mechanism that would motivate platform owners to collude with the vendors by letting the ratings become inflated (Aziz et al. 2020, Filippas et al. 2020, Wang et al. 2020).

Most retail platforms have adopted a familiar 5-star ratings system for displaying evaluations of vendors and their items, which makes it easy to compare average ratings across platforms. Already some time ago, Chevalier and Mayzlin (2006) noted that *"there is nothing to stop a consumer from using the information provided by a one Web site to inform purchases made elsewhere"* (p. 345). Importantly, the average ratings have been found to differ across platforms even for the same vendors (Zervas et al. 2020) or items (Park et al. 2021), which raises a question whether the diverging average ratings could affect the choice of the platform that a consumer chooses to transact with the vendor. On the one hand, one would expect that a platform which provides more accurate product information is able to match supply and demand better but, on the other hand, Sahni (2016) shows that making consumers more informed about company products can be strategically disadvantageous if the benefits accrue to competitors due to spillover effects. However, until now, spillover effects have been studied as a 'horizontal' phenomenon between entities such as products, services, or companies that occupy the same vertical position in the value chain (e.g. Borah and Tellis 2016, Janakiraman et al. 2009, Kumar et al. 2018b, Roehm and Tybout 2006). We hypothesize that the way in which a digital retail platform integrates vendor, their items, and delivery into a unified purchasing experience can give rise to a 'vertical' spillover effect from vendor ratings to platform choice.

The vertical spillover effect implies that consumers prefer to order from a platform where a vendor or an item is rated higher even if they will get the same item regardless of the platform choice. This may happen when a consumer checks a product or service simultaneously on multiple platforms to learn about it and then purchases it on a platform where the item or its vendor is rated higher, or in a less obvious manner, when a consumer first encounters an item on one platform but foregoes the purchase due to its low rating, and then later sees the item on another platform where it is rated higher and now proceeds with the purchase. Importantly, the existence of a vertical spillover effect does not mean that the overall consumption experience is identical between the platforms (which it probably never is) but, all other things being equal, that the consumers tend to order from a platform where the chosen vendor or item is rated higher. In other words, it is as if consumers would believe that they gain some utility from seeing a larger number for the average rating even if the number does not signal a better-quality product or platform than a smaller number on another platform.

To strategically design and govern their ratings systems, platform owners need to know whether there is a vertical spillover effect from average ratings to platform choice. To study this, we ask the following research question: *Do diverging average ratings for the same vendor across platforms influence the choice of the platform used to purchase from the vendor?* The answer is not obvious and has important managerial implications. Our starting point is that, all other things being equal, the diverging average ratings should not affect the choice of a platform used to buy the product. There may be many other reasons related to, for instance, to pricing, promotions, delivery hours, and consumers' *a priori* preference for a particular platform that may entice to choose one platform over the other. However, controlling for such factors, the diverging average ratings between platforms should not affect the platform consumers choose for conducting the transaction because they can expect to get the same item. By contrast, the possibility of a vertical spillover effect suggests that average vendor ratings can also taint the platform itself and hence affect platform choice. The immediate impact of vertical spillovers from average ratings to platform choice may seem innocuous to the consumers who receive after all the same item. Yet, for platform owners such spillovers can provide an incentive to let the ratings become inflated to avoid losing transactions to competing platforms. Progressive ratings inflation will diminish the value of average ratings as a source of information that enables consumers to make successful purchases, especially if the ratings become compressed toward the maximum value (Aziz et al. 2020, Filippas et al. 2020, Kokkodis 2021).

To study the existence of a vertical spillover effect, 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 with different subject pools: undergraduate students and Amazon mTurk workers. We design the experiments to separate the effect of diverging average ratings value from all platform attributes and possible *a priori* preference for one or the other platform among the subjects. In each experiment, subjects choose simultaneously a restaurant and a platform they want to use to place a delivery order from that restaurant, from two restaurant and two platform choice options. For one of the restaurants, the average rating varies between the platforms whereas for the other restaurant the average rating is the same between the platforms. This allows *ceteris paribus*, that is, all other things being equal type reasoning about the impact of diverging average ratings values on platform choice.

For the platforms, we choose Yelp and GrubHub. These two platforms share the same delivery system, and therefore the subjects should expect to get exactly the same food and delivery experience regardless of the platform they choose. In other words, subjects should realize that the difference in average ratings of the same restaurant between two platforms does not signal a difference in the food or the delivery service, while we control for platform attributes and *a priori* preference for one of the platforms through our experiment design. This allows us to isolate the impact of diverging average rating value on platform choice and, hence, to identify the presence of a vertical spillover effect. We also conduct a supplementary study to look for the possible mediation effect of the number of ratings, since a common reason for diverging average ratings between platforms is random variation due to a small number of ratings.

The results show that diverging average ratings between platforms impact platform choice. The subjects prefer to order from the platform where the chosen restaurant is rated higher even when it should be obvious that food quality and delivery experience are the same regardless of the platform they choose. This is explained by a vertical spillover mechanism and suggests that it may be disadvantageous for a platform to counter ratings inflation unless competing platforms are forced to do the same. Consequently, platform owners are required to balance between maintaining the integrity of ratings and trying to avoid strategic mistakes in competition with other platforms. The results have important managerial implications and suggest that a regulatory intervention or restructuring of some sort will be needed to maintain the integrity of the standard 5-star ratings system in the long run. For instance, it may be necessary to separate rating systems from retail platforms to a third party that is not positioned to benefit from ratings inflation. Also, contrary to some earlier findings, we find no evidence of a moderation effect by the number of ratings in our setting. 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 way our study combines a stated preferences method and incentive aligned experiments allows us to discuss the relative merits of the two methods for studying consumer reactions to online ratings. The stated preferences method can often be used where experiments involving the observation of real behavior are not feasible, yet there is a concern that the preferences expressed by survey respondents may differ from their real behavior. To this end, there is evidence that subjects perceive hypothetical and real choices differently. For example, Chartrand et al. (2008, study 3) found that priming participants with real versus hypothetical choices result in significantly different subsequent behavior. Consequently, it is recommended 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). We find that the results of the stated preferences method are consistent with the incentive aligned experiments, although the data from the former are somewhat noisier. The stated preferences method would therefore seem applicable for studying consumer reactions to the 5-star ratings system.

Diverging Ratings Across Platforms

The valence, volume, and variability of ratings and reviews from fellow consumers have a major impact on people's purchase decisions in the online environment (Babic Rosario et al. 2016, Blal and Sturman 2014, Floyd et al. 2014, Forman et al. 2008, Li 2018, Sun 2012, Watson et al. 2018, You et al. 2015, Zimmermann et al. 2018). Studies show that average ratings can often reflect the perceived quality of vendors and their items well (de Matos et al. 2016, Gao et al. 2015, Siering and Janze 2019) but also that the ratings can become variously biased (Hu et al. 2009, Nosko and Tadelis 2015, Saifee et al. 2020) and that consumers often have difficulties in interpreting ratings correctly (de Langhe et al. 2016, Yin et al. 2016). Furthermore, recent studies reveal that the same vendor or item is often rated differently across platforms despite the same objective quality: for instance, Zervas et al. (2020) show that cross-listed properties tend to be rated higher on AirBnB than on TripAdvisor, and Park et al. (2021) compare products sold by both Amazon and BestBuy finding that these are rated differently across platforms. Such differences can result from a combination of factors including heterogeneous consumer tastes (Zimmermann et al. 2018), fake or strategic reviewing (Ho et al. 2017, Kumar et al. 2018b, 2019, Sahoo et al. 2018, Wang et al. 2020), managerial interventions (Ananthakrishnan et al. 2020, Kumar et al. 2018a), and sequential and temporal dynamics (Godes and Silva 2012, Lee et al. 2015, Moe and Schweidel 2012, Park et al. 2021) that affect the body of ratings differently on different platforms.

Among different types of biases that affect ratings, ratings inflation has recently received increasing attention as the phenomenon can threaten the integrity of an entire ratings system (Filippas et al. 2020). Ratings inflation means that average ratings increase over time and lose their variance as they become compressed toward the maximum value (Aziz et al. 2020, Hu et al. 2009, Kokkodis 2021). In the short term, this can increase platform revenues but, in the long term, progressive ratings inflation will render average ratings less informative as the system loses capacity to signal relevant differences between the vendors and their items. Literature identifies different reasons for ratings inflation that relate to reviewer and vendor behavior that 'push' average ratings 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, that is, why platform owners may want to let the ratings become inflated. This is the gap we address by assessing the impact of diverging average ratings (across platforms) on platform choice.

Platform Competition When Both Sides Multihome

Retail platforms increasingly operate in markets where there are multiple strong contenders maneuvering for competitive advantage (Huotari et al. 2017, Rietveld and Schilling 2020, Schilling 2002). Food delivery, ride sharing, and lodging are just a few examples of industries where platform owners need to pay close attention to what competing competitors do as vendors and consumers are typically present on multiple platforms (Bakos and Halaburda 2020, Kim et al. 2017). Under these circumstances, when both sides of the platform multihome, consumers can use alternative platforms to transact with any particular vendor. This makes platform owner's opportunities to capture value from platform interactions precarious as the platform loses its position as the sole gatekeeper between the parties. However, whether the average ratings may affect platform choice (competition) under these circumstances remains an open question—we find only a few studies on the impact of ratings and reviews on platform competition.

Chevalier and Mayzlin (2006) analyze the impact of diverging average ratings on sales and find that a higher average rating for a book on a platform results in more sales as compared to a competing platform that shows a lower average 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, but 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). Kathuria and Lai (2018) argue that the body of ratings and reviews is a strategic asset in platform competition and discuss their portability across platforms from a legal perspective. The authors point out 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). This gets some indirect support from studies showing the impact of review volume on sales (Babic Rosario et al. 2016, Floyd et al. 2014, You et al. 2015), while 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).

However, extant studies have not identified a mechanism by which platform-specific bodies of ratings would influence platform competition and, thus, whether competition could motivate platform owners to let ratings become inflated. We aim to fill this gap in the literature by theorizing and empirically identifying a vertical spillover effect from diverging average ratings to platform choice. This is an important shortcoming in the literature for the presence of a vertical spillover mechanism could motivate platform owners to let ratings become inflated and, hence, threaten the integrity of the standard 5-star system.

Spillover Effects

Spillover effects are defined as externalities by which an event in one context influences another event in a proximate but essentially unrelated context. They can result from different underlying mechanisms (Liang et al. 2019). For instance, a product recall due to a defective component taints perceptions of the recalled product, but they have also been found to have a negative halo on the perceptions of other similar products even if these do not use the defective component (Borah and Tellis 2016). In the latter case, the transfer of a negative evaluation from a product-context to another is unwarranted but nevertheless affects the sales of similar products. Beyond such a transfer of factual evaluations between seemingly similar things, Xu and Schwarz (2018) explain spillover effects with behavioral mind-sets that consumers transfer between contexts, that is, *"spillover effects of mind-sets are expected when the procedures activated by previous goal-directed activities facilitate goal pursuit on a subsequent task"* (p. 52). Thus, it is not always the actual evaluation but the comparative behavioral disposition that consumers carry over to another context. The latter can notably also explain spillover effects between different types of things.

IS and marketing research have analyzed spillover effects from consumer perceptions of a product, service, or a company to the sales of other entities in the same category (e.g., Borah and Tellis 2016, Janakiraman et al. 2009, Roehm and Tybout 2006, Sahni 2016). Spillovers have also been found in the context of online ratings and reviews. 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 use 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. In all these cases, spillovers happen horizontally between entities that occupy the same position in the value chain and could thus be explained as a halo from a thing to another in the same category. In this paper, we extend the idea of spillover effects to settings where consumer evaluations may be transferred between different types of entities.

The way in which retail platforms integrate different parts of the value chain into a unified consumer experience and thus 'invert' the firm (Parker et al. 2017) can give rise to vertical 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 vendor or item across platforms affect platform choice (competition). Li and Agarwal (2017) suggest the presence of such spillovers that happen along the value chain, that is, from a first-party app to third party apps, whereas Liang et al. (2019) find that editorial recommendations of mobile apps on one platform spill over to the sales of the same app on different platforms. Yet, we find no studies that would directly address potential spillovers from the evaluations of platform vendors or their items to the platform itself. The possibility of such vertical spillovers is theoretically interesting and may affect how retail platform owners should govern their ratings. We find it intuitive that the average ratings of the vendors and their items listed on a retail platform could taint perceptions of the platform itself—consumers are known to be susceptible to cognitive biases when interpreting ratings (de Langhe et al. 2016, Katsamakas and Madany 2019, Nosko and Tadelis 2015) and, in a competitive environment, this could affect their platform choice.

In our empirical setting, the presence of a vertical spillover effect would mean that consumers are not only more likely to buy from a vendor that have a higher average rating—this is well known by the literature—but that they are also more likely to buy from a platform where the vendor is rated higher as compared to a competing platform where the same vendor has a lower average rating. The vertical spillover effect can thus result in demand shifting from a platform to another either by a direct platform choice or indirectly as we have discussed above. 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 usually bundled together and sold by a platform such as Waiter.com, Yelp or, again, GrubHub which may collect their own ratings or source these from another platform.

We specifically assume that a vertical spillover effect emerges as follows. To order a meal, a consumer must choose from which restaurant they want to order. This activates a behavioral mind-set that makes comparative procedures and the signal from average vendor ratings cognitively salient to the consumer (Xu and Schwarz 2018). If the consumer then retains the same comparative mind-set with respect to a platform choice but uses (now) a false signal from diverging average vendor ratings across platforms as an indication of a quality difference between the platforms, then vendor ratings spill over to platform choice.

Methodology

We study platform choice in an online food delivery setting using a similar approach to Wu et al. (2021) who deploy a combination of stated preferences and incentive aligned experiments to study ratings. In ‘incentive aligned’ experiments the subjects are motivated to choose according their true preferences, 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 suitable for incentive aligned experiments because i) it allows to set a uniform monetary value for rewards that the subjects receive, ii) it provides the subjects with a flexibility to choose goods, that is, rewards that are likely to be desirable for them, and iii) we can ensure that the subjects cannot transfer their rewards or return them to the vendor in exchange for cash.

In our design we particularly focus on the realism of decision variables: we attempt to make sure that the subjects trust the information we provide them, and that they are confident that their choices affect their rewards in a way we specify in the instructions, while we are less concerned about the realism of the interface in which the choice is made. Because of this, we present the subjects with true information about the local restaurants which they can choose for the food order, but we display this information in a structured survey interface which allows us to control the experimental environment by limiting visible platform attributes to a minimum and by using randomization to control for remaining nuisance factors.

We choose Yelp and GrubHub as the platforms. These are two well-known competitors in the food delivery market in the US. Both platforms use the same GrubHub delivery service, which means that consumers receive the same meal and delivery experience regardless of 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 across the platforms, while the other has the same average rating on both platforms.[1] The restaurant with consistent average ratings across platforms serves as a control: it allows us to control for *a priori* preference for one of the platforms among the subjects. Adding the control restaurant also addresses a possible experimenter’s demand effect by obscuring from the subjects that we are interested in their choice between the platforms (Camerer 2015). We refer to the two restaurants as the *Divergent Ratings Restaurant* and the *Control Restaurant*.

To separate the effect of average ratings from other factors present in a platform user interface, we present the subjects with minimal information needed to make an informed decision, which is retrieved from Yelp and GrubHub restaurant pages: a restaurant cuisine description, average rating, and delivery hours (subjects are allowed to schedule a delivery for a later time during the day). This also eliminates potential

[1] In the sample of 223 restaurants collected for the area where our incentive aligned experiments took place, 207 restaurants are rated on both Yelp and GrubHub. The average rating diverges by at least 0.5 points for 160 restaurants (77.3%) between the platforms.

problems associated with some the nuisance factors present in the user interface changing over the duration of the experiment. To avoid the branding or prior knowledge of the restaurants influencing the results, we label the restaurant choices (rows) as *Restaurant 1* and *Restaurant 2* in the data collection instrument implemented using Qualtrics platform (see Appendix). In the incentive aligned studies, the actual restaurant names are revealed after the subjects have made their choice and completed a demographic survey. In the stated preferences studies, we keep the setting as close 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 restaurants using either Yelp or GrubHub.

Treatments

We implement a between-subject design in which each subject faces only one combination of treatment variables and makes exactly one decision. This approach avoids potential issues due to natural anchoring that can happen when a subject makes subsequent choices (Charness et al. 2012), which are atypical in the current context as consumers do not usually place multiple food orders immediately one after another. The treatment (focus) variable is the average rating for *Restaurant 1* and *Restaurant 2* on Yelp and GrubHub. Next, we discuss these together with our randomization approach to deal with nuisance variables.

Focus Variable. The focus variable of our experimental design is the average restaurant rating in the 5-star system, which is a straightforward one-dimensional metric that most platforms display prominently to summarize consumer evaluations of vendors or their items. We do not consider the content of reviews or the distribution of the individual ratings. Neither Yelp nor GrubHub shows ratings distribution to consumers and, while review content can affect consumer choice, their effect can be difficult to interpret and generalize as they are often subject to nontrivial interaction effects (Vana and Lambrecht 2021, Cho et al. 2021). Most importantly, such factors are secondary from the perspective of our aim to identify a vertical spillover effect from average restaurant ratings to platform choice. However, we perform an additional study to check for the effect of the number of individual ratings upon which the rating average is calculated.

We choose the restaurants so that the *Control Restaurant* has the same average rating on both platforms, whereas the *Divergent Ratings Restaurant* is rated higher than the *Control Restaurant* on one platform and lower than the *Control Restaurant* on the other platform. We refer to the platform where the *Divergent Ratings Restaurant* has a higher average rating as the *High Rating Platform* and the platform where the restaurant has a lower average rating as the *Low Rating Platform*. To control for the possible interaction between the platform and the average rating variables, we include two possible conditions: the *High Rating Platform* can be either Yelp (*YelpHigh* treatments) or GrubHub (*GrubHubHigh* treatments). Note, again, that the quality of the food and the delivery will be the same for the same restaurant regardless of platform choice. We emphasize this on the decision screen and, if the subjects act accordingly, the platform choice should be independent of the choice of the restaurant indicating no spillover effect from restaurant choice to platform choice. By contrast, if we find that platform choice is not independent of restaurant choice, this could mean that the average restaurant ratings spill over to the platform choice.

In the incentive aligned experiments, we need to use information for actual restaurants that deliver to the location where the experiment is conducted. In *GrubHubHigh* condition, the *Divergent Ratings Restaurant* is *Pizza & Grill (PG)* restaurant. The restaurant has an average rating 3.5 on Yelp and 4.5 on GrubHub. The *Control Restaurant* is *Coffee & Pizza (CP)* restaurant that has an average rating 4.0 on both platforms. In *YelpHigh* condition, the *Divergent Ratings Restaurant* is *Middle Eastern (ME)* restaurant. The restaurant has an average rating 4.5 on Yelp and 3.5 on GrubHub. The *Control Restaurant* is *Sandwiches (SW)* restaurant that has an average rating 4.0 on both platforms. Note that *YelpHigh* and *GrubHubHigh* treatments are symmetric in a sense that they have 4.5 and 3.5 average ratings for the *Divergent Ratings Restaurant* and 4.0 average rating for the *Control Restaurant* on both platforms. We also perform an additional incentive aligned experiment to verify the robustness of our findings. In this study, the combination of average ratings is slightly different: the *Divergent Ratings Restaurant* is El *Rincon Latino (RL)* that has an average rating 5.0 on Yelp and 4.0 on GrubHub. The *Control Restaurant* is *Calle Del Sabor (CS)* that has an average rating 4.5 on both Yelp and GrubHub. We thus call the experiment *YelpHigh2*. To summarize, in the incentive alignment study we have three treatments: *GrubHubHigh*, *YelpHigh*, and *YelpHigh2*, which we implement by using three different combinations of local restaurants.

In the stated preferences studies, we are not tied to real restaurant information, but we 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 (ME)* and *Sandwiches (SW)* restaurant pair from the incentive aligned experiments, but we randomly reassign the average ratings so that for some of the subjects the *High Rating Platform* is Yelp, and for the others it is GrubHub. Just like in the incentive aligned treatments *YelpHigh* and *GrubHubHigh*, we have 4.5 and 3.5 average ratings for the *Divergent Ratings Restaurant* and 4.0 average rating for the *Control Restaurant* on both platforms. The random assignment of average ratings to the restaurant–platform combination allows us to control for a possible interaction between the restaurant type and the ratings difference, since the restaurants are equally likely to have diverging ratings in the stated preferences study.

Since the average 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 average 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&G* treatment label means that ME restaurant has 4.5 average rating on GrubHub and 3.5 average rating on Yelp, while the SW restaurant has 4.0 average rating on both platforms. In the next section, we provide details on how we handle nuisance variables by randomization.

Randomization. We are interested in whether the diverging average ratings for the same restaurant across platforms can affect platform choice even if the difference does not suggest a difference in the quality of the purchased meal or its delivery. Therefore, the experiments are designed so that the quality of the choice options remains constant regardless of the value of our focus variable, that is, the average rating associated with the restaurant on a specific platform. However, we also need to control for a few other factors beyond the focus variable that can affect the choice of the restaurant–platform pair. To begin with, the subjects may have *a priori* preference for a particular platform. We account for such preferences by observing the conditional probabilities of choosing one platform over another for those subjects who choose the *Control Restaurant* and using these as the baseline for comparing the conditional probabilities for those subjects who choose the *Divergent Ratings Restaurant*. In other words, we observe whether subject choices shift toward the *High Rating Platform* when they choose the *Divergent Ratings Restaurant* as compared to when they choose the *Control Restaurant*.

Furthermore, we note that the order in which the restaurants and the platforms are presented in the layout of the data collection instrument, slightly differing restaurant category descriptions between the two platforms, and that Yelp stops accepting delivery orders 15 minutes earlier than GrubHub could influence subjects' choices. To avoid picking up effects from such nuisance factors, we randomly shuffle these between the treatments as follows. In the incentive aligned study, we vary the order of the restaurants (both restaurants are equally likely to be labelled as 'Restaurant 1' or 'Restaurant 2') and the order of the platforms (both platforms are equally likely to be presented in the left or the right column of the table) without compromising the realism of the treatments. In the stated preferences study, we perform a comprehensive randomization by further varying the following factors. We randomly swap the platform on which the *Divergent Ratings Restaurant* is rated higher, the category descriptions for each of the two restaurants (four possible combinations), and the delivery hours (either Yelp or GrubHub closes 15 minutes earlier) between the platforms. We implement the randomization by assigning a random treatment to each subject in real time to eliminate any systematic composition of subjects to a specific treatment.

Participants

The study subjects are recruited from three different pools. For the experiments based on stated preferences, we use i) undergraduate business majors in a Northern American university who were offered a course credit for participating in the study during Fall 2019 semester, and ii) US-located Amazon Mechanical Turk (mTurk) workers who received $2 reward for completing the survey in June 2020. These two pools of subjects participated in the stated preferences experiments online from their own private location. For the incentive aligned experiments, we iii) 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 treatments *YelpHigh* and *GubHubHigh* were conducted in person during Fall 2019 and Spring 2020 semesters, and *YelpHigh2* treatment was conducted in Fall 2021 via Zoom, due to the restrictions imposed by the global pandemic. Zoom sessions were identical to the in-person sessions, except that a participant had to forward the menu selections and the delivery address to the researcher via Zoom chat, and the researcher then placed that order and made a payment. The subjects in the incentive aligned study received no other reward than the food order that they placed during the study.

Study 1

There is a total of 301 undergraduate students who participated in our first stated preferences study. The upper row in Figure 1 summarizes the subjects' choices for each of the four treatments. To begin with, we observe that the subjects have clear *a priori* preference for GrubHub over Yelp—the subjects who choose the *Control Restaurant* prefer to place the order through GrubHub. We compare the proportion of subjects who choose GrubHub over Yelp for *Control Restaurant* and *Divergent Ratings Restaurant* to assess whether the platform choice is independent from the restaurant choice. Recall that under our null hypothesis the platform choice is independent of the restaurant choice indicating no spillover effect. If this would be true, the relative heights of the paired yellow and green bars would the same or very similar in each tab. This is clearly not the case. The subjects' platform choice is dependent of their restaurant choice and, importantly, the direction of the differences consistently shows that the subjects who choose *Divergent Ratings Restaurant* prefer to place their order through a platform where the restaurant is rated higher. This suggests a presence of a spillover effect from the average restaurant rating to the platform choice. We further assess whether the differences in proportions are statistically significant by chi-squared tests and find that all treatments are significant at least on 10% level (p-value for each treatment is shown in the respective tab in Figure 1).

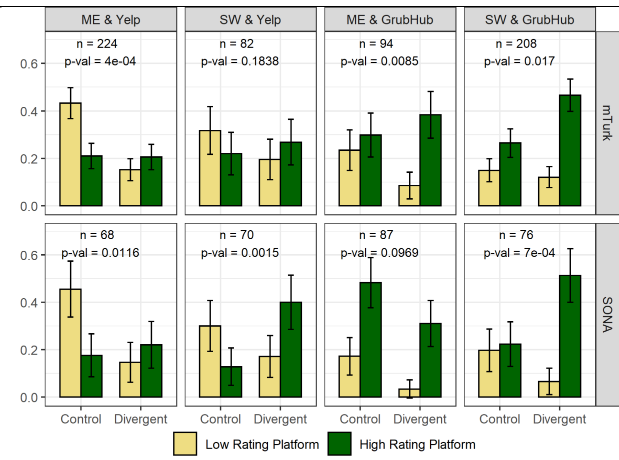

Figure 1. Subject choices in stated preferences studies 1 and 3

Study 2

In the second study, we shift to an incentive aligned experiment to assess whether the results hold when the subjects' incentives are aligned with the study aims. The experiments are considerably more laborious to perform as compared to the stated preferences method and therefore we have fewer subjects in the experiments. However, since the subjects' incentives should now be better aligned with the aims of the experiment, we should be able to observe a significant deviation from the null hypothesis in a much smaller sample. We assign 13 subjects to *YelpHigh* treatment and 16 subjects to *GrubHubHigh* treatments in an in-person lab setting. The left and middle panels in Table 1 show the number of subjects who chose each option in each treatment in Study 2, with the average 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 (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 suggest that the platform choice is dependent on the restaurant choice, which is consistent with the presence of a vertical spillover effect.

To test the independence of the platform choice from the restaurant choice statistically, we use Pearson's Chi-squared test with p-values obtained by Monte Carlo simulation with 106 replicates which is an appropriate approach for small samples (Lin et al. 2015). The results are significant at 5% level for *YelpHigh* (p-value 0.0210) treatment but not significant for *GrubHubHigh* (p-value ≈ 1) treatment. Note that it is considerably harder to detect a significant difference in the *GrubHubHigh* treatment due to strong *a priori* preference for GrubHub among the subjects—we have only two subjects who chose the *Control Restaurant* and the Yelp platform. However, the direction of the effect in *GrubHubHigh* treatment is similar to the other treatment and the findings from Study 1, and thus consistent with a vertical spillover effect.

	$YelpHigh$			$GrubHubHigh$			$YelpHigh2$	
	Yelp	GrubHub		Yelp	GrubHub		Yelp	GrubHub
ME	5 (⋆ 4.5)	2 (⋆ 3.5)	P&G	1 (⋆3.5)	10 (⋆4.5)	RL	5 (⋆ 5)	6 (⋆ 4)
SW	0 (⋆ 4.0)	6 (⋆ 4.0)	C&P	1 (⋆4.0)	4 (⋆4.0)	CS	1 (⋆ 4.5)	14 (⋆ 4.5)

Table 1. Subject choices in incentive aligned studies 2 and 4

Study 3

In the third study, we again use a stated preferences method but with a different subject pool to account for the possibility of a certain type of campus subjects or the locations driving the results, and to extend our results regarding the number of ratings on which the average ratings is based. We use 608 mTurk workers based in the US who receive a $2 reward for participating in our study.

First, we replicate our results about platform choice in the new subject pool using a different type of reward. The lower row in Figure 1 summarizes the subjects' choices for each of the four treatments. The results are very similar to Study 1 and 2. We again observe that the subjects have *a priori* preference for GrubHub over Yelp and that the subjects' platform choice is dependent of their restaurant choice. This is confirmed by chi-squared tests showing that all but one treatment is significant on 5% level. The direction of the differences consistently shows that the subjects who choose *Divergent Ratings Restaurant* prefer to place their orders through a platform where the restaurant is rated higher. Therefore, despite different methods, subject pools, and rewards in the three studies, the choice patterns are similarly consistent with the vertical spillover effect.

Second, until now we have excluded as many nuisance factors as possible to isolate the effect of average ratings (valence) on platform choice. These include the number of ratings (volume) that have been shown to affect consumer behavior. Yet, when choosing the restaurants for our study we observed that diverging ratings across platforms often occur when the restaurant is new to the platform and it has not yet had time to accumulate many ratings. Diverging average ratings across platforms could thus be an artifact of variation due to a small sample. If the subjects perceive that the sample is too small to provide a reliable estimate of the restaurant quality, 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?"*

Our focus variable is the number of ratings n that takes one of the values: 1, 3, or 15. To observe if the direction of the effect of the number of ratings is different for high and low rated restaurants, we have *Low Rating Condition* where the average rating r is 2.0, and *High Rating Condition* where the average rating r is 4.0. To ensure that a sufficient percentage of subjects select each option, we adjust the corresponding values of the gift card *g* to be \$50 in *Low Rating Condition* and \$20 in *High Rating Condition*. The stated values of the gift cards and cash were established based on a pilot study so that a reasonable proportion of subjects would choose cash versus a gift card. In this part of the study, we implement a within-subject design, where some of the subjects face the *Low Rating Condition* first, while others face the *High Rating Condition* first. In each of the two conditions the number of ratings n is drawn independently from the three possible values. Thus, we have a full factorial design (2 × 3) with two average rating conditions combined with three conditions based on a different number of ratings.

We recruit a total of 491 subjects 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 treatment. 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. Also, note that in *Low Rating Condition* we would not expect the number of subjects who choose the gift card to increase as the average rating becomes a more reliable signal of low quality.

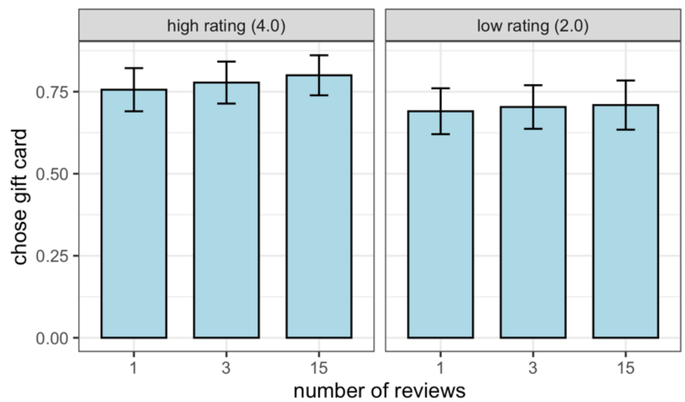

Figure 2. The impact of the number of ratings on the choice between a gift card and a cash reward

If differences exist, they are very small: comparing the treatments with 1 and 15 ratings, we find that the 95% confidence interval between the proportion of subjects who chose the gift card and cash is between -0.084 and 0.121 in *Low Average Rating* condition and between -0.046 and 0.136 in *High Average Rating* condition. Overall, the number of ratings appears to have very little impact on the probability of subjects choosing the restaurant gift card rather than opting out and taking the cash. We conclude that the number of ratings does not appear to mediate the impact of average ratings in the current setting. The fact that diverging ratings across platforms can have an impact on platform choice even if the average rating is based on just a few ratings is concerning, as it suggests that the vertical spillover effect could make it harder for new platforms to enter the market as consumers do not give them the benefit of doubt against incumbents who have already established high valence and volume of ratings.

Study 4

In the fourth study, we want to further increase the robustness of our findings using a different set of ratings values. We recruited 26 university staff and students and assigned them to *YelpHigh2* treatment in online lab sessions to test if the result hold using a different set of average rating values. The right panel in Table 1 shows the number of subjects who chose each option in Study 4, with the average rating for the corresponding platform–restaurant combination shown in parentheses. Again, the results show that the platform choice is dependent on the restaurant choice as the subjects who choose *Divergent Ratings* Restaurant prefer to place their orders through a platform where the restaurant is rated higher, which is consistent with the presence of a vertical spillover effect. Similarly to Study 2, we use Pearson's Chi-squared test with p-values obtained by Monte Carlo simulation with 106 replicates and find that the results are significant at 10% level (p-value 0.0545). Note that the study was conducted in Fall 2021 showing that the vertical spillover effect can still be observed regardless of any changes to the consumer online ordering habits due to the global pandemic.

Discussion and Conclusions

It is well known that the standard 5-star ratings system has a major impact on consumer behavior and it has been suggested that the ratings may also have an impact on the competition between platforms (e.g., Chevalier and Mayzlin 2006, Kathuria and Lai 2018). However, no study has until now identified a mechanism by which the 5-star system would directly influence platform competition. Our study fills this gap by focusing on the impact of diverging average ratings across platforms on platform choice. We conduct a series of experiments using different methods and subject pools showing that consumers are more likely to transact with a particular vendor on a platform where the vendor is rated higher, even if the diverging ratings may not signal any difference in the overall consumption experience. We explain this by a vertical spillover effect as the evaluations of vendors spill over to perceptions about the quality of competing platforms. The results have important managerial and regulatory implications, and allow us to make contributions to literature on spillover effects, platform competition, and online ratings. Finally, we reflect upon methodological differences between the incentive aligned and stated preferences studies in the context of online ratings.

Managerial Implications

Our results show that a threat from hosting lower average ratings than competitors is real when both sides of the platform multihome. Consumers prefer to transact on a platform where the chosen vendor or item is rated higher and, consequently, platform owners need to manage their ratings systems strategically. First, whenever both vendors and consumers multihome on competing platforms, a platform owner needs to balance between safeguarding the integrity of ratings and not ending up providing a virtual showroom to competitors. The platform owner must keep on an eye how competing platforms govern their ratings and react if it seems that a competitor adopts a substantially more relaxed approach to dealing with attempts to artificially boost ratings. For instance, a platform owner may want to follow a sample of vendors on its own and competitors' platforms to see how their average ratings evolve. If the valence and volume of ratings and reviews in the sample evolves differently on the platforms, the platform owner may need to let the ratings become inflated along with the competition or, perhaps, consider adopting an alternative ratings system to avoid the competitive disadvantage from the vertical spillover effect.

Second, research shows that multidimensional ratings systems offer some advantages over the standard 5-star system (Chen et al. 2018, Kokkodis 2021, Schneider et al. 2021, Tunc et al.), and suggests that under certain conditions it may not be beneficial to a retailer to implement a word-of-mouth system at all (Huang et al. 2019). Platform owners could also try different ratings cardinalities or base their word-of-mouth system more on the qualitative reviews (Jiang and Guo 2015, Lee and Kai 2020). Adopting an alternative to the standard 5-star system is a feasible option especially to new entrants whose vendors do not yet have established their ratings on the platform.

Third, while increasing competition generally benefits consumers, our results show that competition between platforms that use the standard 5-star system can drive platform owners to govern their ratings in a way that results in ratings inflation. The risk is that inflated ratings negatively affect consumers' capacity to make successful purchases and, consequently, erode trust in the 5-star system in general. The effects of ratings inflation can already be seen on some major platforms such as AirBnB (Zervas et al. 2020), eBay (Nosko and Tadelis 2015), Uber, and on certain online labor platforms (Kokkodis 2021), where average ratings have become compressed toward the maximum of five stars. While one cannot expect an individual platform to voluntarily put in place mechanisms to curb ratings inflation if that means that the platform is going to lose out in the competition, a regulator could, at least in principle, consider enforcing the separation of the provision of ratings and reviews from retail platform business. This would not only remove direct incentives to let the ratings become inflated but also make it easier for new entrants to enter the market if they could tap into the same pool of ratings and reviews that the incumbents use.

Fourth, we observe that the number of ratings does not seem to mitigate the impact of diverging average ratings across platforms, which adds to the mixed results about the impact of the number of ratings on consumer behavior (Blal and Sturman 2014, Watson et al. 2018, Zimmermann et al. 2018). This is especially concerning in a case of new vendor that enters a platform, which, despite offering quality products and services, may get a low first rating due to randomness. To increase the chances of success for new businesses, platforms may consider hiding the average rating until enough ratings has been accumulated. For example, the restaurant booking platform OpenTable is currently using this approach. The over-reaction to average ratings valence based on small samples is also concerning for new platforms. If consumers are not giving the benefit of doubt to a new platform with fewer ratings than incumbent platforms with vendor ratings already compressed toward the maximum, then the entrant whose average vendor ratings will likely vary more is at disadvantage against more established competitors. Such problems could again be mitigated by using a third-party ratings provider.

Theoretical and Methodological Contributions

In information systems and marketing literature, spillover effects are usually studied as product, company, or brand-related events that influence proximate entities occupying the same vertical position in the value chain. We add to these findings by identifying a spillover effect that takes place between entities occupying different vertical positions in the value chain. In our setting, the vertical spillover effect results from consumers transferring a comparative mindset activated in a vendor selection context to a platform choice context and concomitantly misinterpreting diverging average ratings for the same vendor between platforms as a signal about relative platform quality. It is intuitive that consumers are not always able to attribute quality signals from average ratings sharply to the assessed entity and that the perceptions of vendor quality can have a halo on the surrounding platform context (Borah and Tellis 2016). What makes the vertical spillover effect theoretically important is that it affects the competition between platforms and, consequently, feeds back to the ratings in the form of ratings inflation.

Previous studies suggest that online ratings can become a strategic asset in platform competition (Chevalier and Mayzlin 2006, Kathuria and Lai 2018), and that the ratings tend to become inflated over time (Athey et al. 2019, Filippas et al. 2020, Hu et al. 2009, Kokkodis 2021, Nosko and Tadelis 2015, Zervas et al. 2020). Our results reveal an important connection between these two phenomena by showing that a platform that has higher average vendor ratings can 'steal' transactions from a competitor that has lower average ratings for the same vendors. This is due to a vertical spillover effect. Accordingly, a platform that has lower but perhaps more accurate average ratings may end up suffering from virtual showrooming if consumers use information provided by the platform for choosing from which vendor to buy but then conduct the transaction on another platform. Importantly, our experimental design controls for the platform attributes that are part of the treatment as nuisance variables, and for *a priori* preference for one platform over another among our subjects. This means that we do not make an unrealistic assumption that different platforms would offer an identical consumption experience but instead show that consumers are, all other things being equal, more likely to choose to buy from a platform where a vendor has a higher average rating number. Providing more accurate but lower ratings can therefore be disadvantageous to a retail platform when both vendors and consumers multihome.

Literature identifies vendor manipulation as one of the main reasons for ratings inflation (Fradkin et al. 2015, Hu et al. 2009, Lee and Kai 2020). It is obvious why individual vendors are interested in boosting their own ratings, but the literature has not explained what motivates platform owners to let the ratings become inflated. We fill this gap by showing that a platform can gain competitive advantage from higher average vendor ratings due to a vertical spillover effect, which incentivizes the platform owner to tolerate behavior that contributes to ratings inflation. The results extend the findings about the capacity of positive fake reviews to boost platform sales in the short term (Wang et al. 2020) by showing the strategic nature of ratings inflation when both sides multihome. If combating ratings manipulation is costly (as one would expect), then a platform that strives to offer accurate ratings may effectively invest in operations that hurt itself in competition with those who adopt a more relaxed policy toward attempts to artificially boost the ratings. The average ratings are particularly prone to become inflated whenever both vendors and consumers multihome, eroding the usefulness of the 5-star ratings system in the long run.

Finally, the way our study combines a stated preferences method and incentive aligned experiments allows us to make further remarks about the relative merits of the two methods for studying consumer reactions the 5-star system. The former method is widely used in marketing and information systems research as well as in the industry, where 'would you' type surveys often provide input for important marketing and design decisions. By contrast, there are several reasons why it is not always possible to conduct incentive aligned experiments. In our study, the stated preferences method yields results that are consistent with the incentive aligned experiments with respect to the platform choice, although, as one would expect, the data is less noisy in the incentive aligned experiment compared to the stated preferences experiments. The 'noisier' subject choices imply that the stated preferences method requires many more subjects, but even as such it is often significantly less costly and more flexible to implement than the incentive aligned method.

Conclusion

We have shown that diverging average ratings for the same vendor across competing platforms can trigger a vertical spillover effect that influences platform choice when both side of a retail platform multihome. The vertical spillover effect can give a competitive advantage to a platform with higher average ratings, which motivates platform owners to let the ratings become inflated. The results indicate that platforms need to govern their ratings systems strategically and that increasing competition between platforms may threaten the integrity of the standard 5-star ratings system as a reliable source of information about product quality. To this end, Filippas et al. (2020, p. 1) speculate that ratings systems *"sow the seeds of their own irrelevance"* by becoming increasingly toothless due to the inflation of average ratings. More studies are needed to fully understand potentially self-defeating dynamics affecting the familiar 5-star system that has become a standard feature of most retail platforms.

References

Ananthakrishnan, U. M., Li, B., and Smith M. D. 2020. A tangled web: should online review portals display fraudulent reviews? *Information Systems Research* (31:3), pp. 950–971.

Athey, S., Castillo, J. C., and Chandar, B. 2019. Service quality in the gig economy: empirical evidence about driving quality at Uber. *SSRN*.

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

Babic Rosario, A., Sotgiu, F., De Valck, K., and Bijmolt T. H. A. 2016. The effect of electronic word of mouth on sales: a meta-analytic review of platform, product, and metric factors. *Journal of Marketing Research* (53), pp. 297–318.

Bakos, Y., Halaburda, H. 2020. Platform competition with multihoming on both sides: subsidize or not? *Management Science* (66:12), pp. 5599–5607.

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

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

Camerer, C. F. 2015. The promise and success of lab–field generalizability in experimental economics: a critical reply to Levitt and List, in *Handbook of Experimental Economic Methodology*, G. R. Fréchette and A. Schotter (eds.), Oxford, UK: Oxford University Press.

Charness, G., Gneezy, U., Kuhn, M. A. 2012. Experimental methods: between-subject and within-subject design. *Journal of Economic Behavior & Organization* (81:1), pp. 1–8.

Chartrand, T.L., Huber, J., Shiv, B., and Tanner, R. J. 2008. Nonconscious goals and consumer choice. *Journal of Consumer Research* (35:2), pp. 189–201.

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

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

Cho, H. S., Sosa, M. E., and Hasija, S. 2021. Reading between the stars: understanding the effects of online customer reviews on product demand. *Manufacturing & Service Operations Management*.

de Langhe, B., Fernbach, P. M., and Lichtenstein, D. R. 2016. Navigating by the stars: investigating the actual and perceived validity of online user ratings. *Journal of Consumer Research* (42), pp. 817–833.

de Matos, M. G., Ferreira, P., Smith, M. D., and Telang, R. 2016. Culling the herd: using real-world randomized experiments to measure social bias with known costly goods. *Management Science* (62:9), pp. 2563–2580.

Filippas, A., Horton, J. J., and Golden, J. 2020. Reputation Inflation.

Floyd, K., Freling, R., Alhoqail, S., Cho, H. Y., and Freling, T. 2014. How online product reviews affect retail sales: a meta-analysis. Journal of Retailing (90:2), pp. 217–232.

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

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

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

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

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

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

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

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

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

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

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

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

Katok, E. 2011. Using laboratory experiments to build better operations management models. *Foundations and Trends in Technology, Information and Operations Management* (51), pp, 1–86.

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

Kim, B., and 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), pp. 36–53.

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

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

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

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

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

Lee, Y. J., Hosanagar, K., and Tan, Y. 2015. 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), pp. 739–754.

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

Liang, C., Shi, Z., and Raghu, T. S. 2019. The spillover of spotlight: platform recommendation in the mobile app market. *Information Systems Research* (30:4), pp. 1296–1318.

Lin, J. J., Chang, C. H., and Pal, N. 2015. A revisit to contingency table and tests of independence: bootstrap is preferred to chi-square approximations as well as Fisher's exact test. *Journal of Biopharmaceutical Statistics* (25:3), pp. 438–458.
McGee, P. 2020. Apple: how app developers manipulate your mood to boost ranking. *Financial Times*, https://www.ft.com/content/217290b2-6ae5-47f5-b1ac-89c6ccebab41.
Moe, W. W., and Schweidel, D. A. 2012. Online product opinions: incidence, evaluation, and evolution. *Marketing Science 31(3)*.
Murphy, R. 2020. Local consumer review survey 2020. Technical report, *BrighLocal*.
Nosko, C., and Tadelis, S. 2015. The limits of reputation in platform markets: an empirical analysis and field experiment. Technical Report 20830, *National Bureau of Economic Research*.
Park, S., Shin, W., and Xie, J. 2021. The fateful first consumer review. *Marketing Science* (40:3), pp. 481–507.
Parker, G., Van Alstyne, M., and Jiang, X. 2017. Platform ecosystems: how developers invert the firm. *MIS Quarterly* (41:1), pp. 255–266.
Rietveld, J., and Schilling, M. A. 2020. Platform competition: a systematic and interdisciplinary review of the literature.
Roehm, M. L., and Tybout, A. M. 2006. When will a brand scandal spill over, and how should competitors respond? *Journal of Marketing Research* (43:3), pp. 366–373.
Sahni, N. S. 2016. Advertising spillovers: evidence from online field experiments and implications for returns on advertising. *Journal of Marketing Research* (53), pp. 459–478.
Sahoo, N., Dellarocas, C., and Srinivasan, S. 2018. The impact of online product reviews on product returns. *Information Systems Research* (29:3), pp. 723–738.
Saifee, D. H., Zheng, Z., Bardhan, I. R., and Lahiri, A. 2020. Are online reviews of physicians reliable indicators of clinical outcomes? A focus on chronic disease management. *Information Systems Research* (31:4), pp. 1282–1300.
Schilling, M. A. 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), pp. 387–398.
Schneider, C., Weinmann, M., Mohr, P. N. C., and vom Brocke, J. 2021. When the stars shine too bright: the influence of multidimensional ratings on online consumer ratings. *Management Science* (67:6), pp. 3871–3898.
Siering, M., and Janze, C. 2019. Information processing on online review platforms. *Journal of Management Information Systems* (36:4), pp. 1347–1377.
Sun, M. 2012. How does the variance of product ratings matter? *Management Science* (58:4), pp. 696–707.
Thomas, B. 2021a. E-commerce & online auctions in the US. Indutry Report 45411A, *IBISWorld*.
Thomas, B. 2021b. Mail order in the US. Industry Report 45411B, *IBISWorld*.
Tunc, M. M., Cavusoglu, H., and Raghunathan, S. (forthcoming) Online product reviews: is a finer-grained rating scheme superior to a coarser one? *MIS Quarterly*.
Turner, E., and Rainie, L. 2020. Most Americans rely on their own research to make big decisions, and that often means online searches. Technical report, *Pew Research Center*.
Vana, P., and Lambrecht, A. 2021. The effect of individual online reviews on purchase likelihood. *Marketing* Science (40:4), pp. 708–730.
Wang, Z., Kumar, S., and Liu, D. 2020. On platform's incentive to filter fake reviews: a game-theoretic model, in *Proceedings 41st International Conference on Information Systems*, Hyderabad, India.
Watson, J., Ghosh, A. P., and Trusov, M. 2018. Swayed by the numbers: the consequences of displaying product review attributes. *Journal of Marketing* (82:6), pp. 109–131.
Wu, J., Zhao, H., and Chen, H. 2021. Coupons or free shipping? Effects of price promotion strategies on online review ratings. *Information Systems Research* (32:2), pp. 633–652.
Xu, A. J., and Schwarz, N. 2018. How one thing leads to another: spillover effects of behavioral mind-sets. *Current Directions in Psychological Science* (27:1), pp. 51–55.
Yin, D., Mitra, S., and 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), pp. 131–144.
You Y, Vadakkepatt GG, Joshi AM (2015) A meta-analysis of electronic word-of-mouth elasticity. Journal of Marketing 79:19–39.
Zervas, G., Proserpio, D., and Byers, J. W. 2020. A first look at online reputation on Airbnb, where every stay is above average. *SSRN*, https://ssrn.com/abstract=2554500.

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

Appendix: Subject's Choice Screen

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