**Submitted to *Information Systems Research***
**manuscript (Please, provide the manuscript number!)**

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

**(Authors' names blinded for peer review)**

Common 5-star ratings system is a key source of information for consumers trying to decide where to eat, what products to buy, or which doctor to visit. Therefore, it is not surprising that platform vendors often want to inflate their ratings to boost sales, but it is less clear why some platform owners let them do so. Biased ratings diminish the capacity of the platform to match supply and demand effectively and, hence, a platform owner should act against such ratings inflation. To explain this important dilemma, we study a market in which both vendors and consumers multihome on several platforms and, consequently, the consumers may see different average ratings for the same vendors across the platforms. Using incentive aligned experiments and experiments based on stated preferences method in an online food ordering context, we show that consumers are more likely to buy from a platform where the chosen vendor is rated higher due to a vertical spillover effect. Our results suggest that platform owners need to take a threat from hosting lower average ratings seriously and, at minimum, keep on an eye how competitors govern ratings on their platforms—a platform that naively invests in countering ratings manipulation risks hurting itself if competitors let their ratings become inflated. Adopting a common third-party ratings provider across competing platforms or using alternative ratings systems could work against ratings inflation and make it easier for new platforms to enter the market. The results contribute to academic literature on spillover effects, platform competition and online ratings.

*Key words*: behavioral experiment, multihoming, platform competition, ratings inflation, vertical 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. The vast majority of 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 average rating (Murphy 2020). 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 and attempts to artificially boost the numbers by platform vendors. The latter can contribute to ratings inflation by which average ratings increase over time threatening to render the standard 5-star ratings system less informative to consumers on many platforms (Filippas et al. 2020, Kokkodis 2021, Zervas et al. 2020).

A retail platform owner should ideally work against various biases including ratings inflation to make sure ratings provide accurate information about vendors and item quality to consumers. However, upward biased ratings and fake reviews can drive sales which means that it is not always clear what the platform owner will or should 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 make an effort to 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 are able to match supply and demand in the long run. 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 offering many of the same vendors and items. Recent studies suggest that if both consumers and vendors multihome, that is, use several platforms simultaneously, then the opportunities of the platform owner to capture value from their interactions become 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). A platform owner has to take into consideration the fact that consumers can always take their transactions to a competing platform in case this seems beneficial to them.

### 1.1. Motivation

It intuitively makes sense that a retail platform owner does not want its vendors to look better on the competitor's platform. A few studies have suggested 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 mobile platforms (McGee 2020). Yet, we do not know if there is a real threat from hosting lower average ratings than the competitor or whether the threat is merely an unfounded perception. This us 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 different countermeasures (Ananthakrishnan et al. 2020, de Langhe et al. 2016). The important unanswered question is whether there is a mechanism that would actually 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 standard 5-star ratings system for displaying evaluations of vendors and their items, which makes it easy for consumers to compare average ratings across platforms. Chevalier and Mayzlin (2006) noted already some time ago 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). Interestingly, the ratings have also 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 ratings could affect the choice of the platform that a consumer chooses to transact with a vendor. The answer is not trivial since, on the one hand, one would expect that a platform that provides more accurate product information should be able to match supply and demand more effectively yet, on the other hand, for instance 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.

Spillover effects have been studied until now mostly as a horizontal phenomenon between entities such as products, service, 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). By contrast, we hypothesize based on the work of Li and Agarwal (2017) that the way in which digital retail platforms integrates vendors, their items, and delivery into a unified overall consumption experience can give rise to a vertical spillover effect that affects platform choice. The vertical spillover effect would imply that consumers prefer to order from a platform where a vendor or item is rated higher even if they will get exactly 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 vertical spillover effects do not entail that the overall consumption experience would be identical between the platforms (which it probably never is), only that consumers have a tendency to order from a platform where the chosen vendor or item is rated higher.

### 1.2. Research Question

In order to strategically design and govern their ratings systems, platform owners need to know whether there can be 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?* Based on the above discussion, the answer is not obvious and has important managerial implications. The diverging average ratings should not affect the platform choice above and beyond *a priori* preference for a particular platform, since the consumer can expect to get the same item regardless of the platform they use for conducting the transaction. However, the possibility of vertical spillover effects suggests that vendor ratings can also taint the platform itself and hence affect platform choice. This would provide the platform owner an incentive to let the ratings become inflated, whereas 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, progressive ratings inflation will diminish the value of ratings as a source of information that enables consumers to make successful purchases if the ratings become compressed toward the maximum value (Aziz et al. 2020, Filippas et al. 2020, Kokkodis 2021).

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. In each experiment, subjects choose simultaneously a restaurant and a platform they want to use to place a delivery order from that restaurant. For some of the restaurants, the average rating varies between platforms whereas for others the average rating is the same between the platforms. We design the experiments so that they control for different platform attributes and, importantly, possible *a priori* preference for one or the other platform among the subjects.

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 must realize that the difference in ratings of the same restaurant between two platforms does not signal the difference in service provided by the platforms themselves. This isolates the impact of diverging average ratings on platform choice and, hence, allows us 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 tend to choose the platform where the chosen restaurant is rated higher even when it should be obvious that food quality and delivery experience will be the same regardless of the platform they choose. This is explained as a vertical spillover effect and suggests that it may be disadvantageous for a platform to counter ratings inflation unless the competing platforms can be made to reciprocate the actions. Platform owners are thus 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 5-star system in the long run. For instance, it may be necessary to separate rating systems from platforms to a third party that is not positioned to benefit from ratings inflation. Also, contrary to some earlier findings (REFs), 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 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).[1] We find that the results of the stated preferences method are remarkably consistent with the incentive aligned experiments, although the data from the former

[1] 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 process 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, while they were just not motivated to carefully consider the setup.

are somewhat noisier than from the latter. The stated preferences method would therefore seem applicable for studying consumer reactions to the familiar 5-star ratings system.

## 2. Literature Review

In this section, we first summarize relevant findings on online ratings and why they may diverge between platforms, and then consider known mechanisms that push ratings to become inflated. Second, we review literature on platform competition and summarize a few available papers on how the 5-star system influences the competition. Each subsection concludes with a distinct gap in the literate that we target with our study.

### 2.1. 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 reflect the perceived quality of vendors and their items remarkably 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). Interestingly, recent studies reveal that the same vendors and items are often rated differently across platforms despite identical 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. The diverging ratings for the same item 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.

Ratings inflation has recently received increasing attention among different types of biases that affect ratings as it can potentially threaten 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). The phenomenon can increase platform sales in the short term, whereas in the long term ratings inflation threatens to render ratings uninformative by diminishing the capacity of the system to signal relevant differences between the vendors or their items. Lee and Kai (2020) find in their literature review four mechanisms driving ratings inflation: social herding, self-selection bias, the under-reporting of negative experiences, corporate manipulation of ratings (see also Fradkin et al. 2015). We note that all these are related to reviewers or vendors and, hence, only account for push toward higher average ratings. However, the mechanisms cannot explain why platform owners may want to let the ratings become inflated. Our study aims to fill this gap in the literature by assessing the impact of average ratings on platform competition.

### 2.2. Platform Competition When Both Sides Multihome

Retail platforms increasingly operate in markets where there are multiple strong contenders manoeuvring 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 competitors as their vendors and consumers are typically present on multiple competing platforms (Bakos and Halaburda 2020, Kim et al. 2017). Under these circumstances, when both sides of the platform multihome, the 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. This is surprising given their substantial impact on consumer purchase decisions.

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 with a lower rating for the same book. The authors do not discuss whether the diverging average ratings affect consumer choice between the platforms, 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), which gets indirect support from by studies showing the impact of review volume on sales (Babic Rosario et al. 2016, Floyd et al. 2014, You et al. 2015, Watson et al. 2018). Despite these insights, neither of the studies delve into mechanisms that could explain how online ratings and reviews affect platform competition. 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.

## 3. Spillover Effects

Spillover effects are 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 can obviously taint the 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 necessarily use the same component (Borah and Tellis 2016). The transfer of negative evaluation from a product-context to another may be unwarranted but it nevertheless affects the sales of other 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 explain spillover effects between different types of things.

Information systems 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 of 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 a setting in which consumer evaluations are transferred between two different types of entities.

We argue that the way in which digital 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. To this end, 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 and 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. However, 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 vendors. We find it intuitive that the average ratings of the vendors and their items listed on a retail platform could also taint the perceptions of the platform itself—consumers are known to be susceptible to cognitive biases when interpreting ratings information (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 higher average ratings, 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 is also present but 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 vertical spillovers may emerge 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 highly salient to the consumer (Xu and Schwarz 2018). If the consumer then retains the same comparative mind-set with respect to a platform choice, as we would expect, but also uses the now false signal from diverging average vendor ratings between platforms from the previous choice context, then the vendor ratings spill over to platform choice.

## 4. 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 online ratings. The food delivery setting is particularly 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 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.

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 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 across the platforms, while the other has the same average rating on both platforms.[2] 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 actually interested in their choice between the platforms (Camerer 2015). We refer to the two restaurants as the *Divergent Ratings Restaurant* and the *Control Restaurant*.

[2] 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.

In order to separate the effect of average ratings from other factors that are 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 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 in Qualtrics (see Appendix). In the incentive aligned studies, the actual restaurant names are revealed only 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 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. This approach avoids potential issues due to natural anchoring that can happen when a subject makes subsequent choices (Charness et al. 2012), which is not present in reality as consumers do not usually place multiple food orders immediately one after another. The treatment (focus) variable is the ratings for *Restaurant 1* and *Restaurant 2* on Yelp and GrubHub. Next, we discuss these together with our randomization approach to deal with nuisance variables.

**4.1.1. 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

**Table 1** **Studies.**

| Study | Method | Subject pool | n | Reward | Setting and time period |
|---|---|---|---|---|---|
| 1 | Stated preferences | Business majors | 301 | Course credit | Online, Fall 2019 |
| 2 | Incentive aligned | Staff & students | 29 | Meal order | In-person lab, Fall 2019—Spring 2020 |
| 3 | Stated preferences | mTurk workers | 608 | $2 | Online, June 2020 |
| 4 | Incentive aligned | Staff & students | 26 | Meal order | Online lab, Fall 2019 |
| 5 | Stated preferences | mTurk workers | | $2 | Online, |

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. By contrast, if we find that the platform choice is not independent of the choice of the restaurant, this means that the 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 close 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.

**4.1.2. Randomization.** We are interested in whether the diverging average ratings for the same vendor across platforms can affect platform choice even if the difference does not suggest a difference in the quality of the purchased item 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 in our experimental design. 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*.

Furthermore, we note that the order in which the restaurants and the platforms are presented in our web form layout, 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 can 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 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, including the potential confounding effect of demographic factors.

### 4.2. Participants

The study subjects are recruited from three different pools. For the stated preferences method experiments, 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.

## 5. Results

The findings consistently show that consumers' platform choice is affected by diverging average vendor ratings across platforms.

### 5.1. 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 Restaurant* in order 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. 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 orders 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-values for each treatment is shown in the respective tab in Figure 1.

### 5.2. 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 run 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 2 show the number of subjects who chose each option in each treatment in Study 2, with the average rating

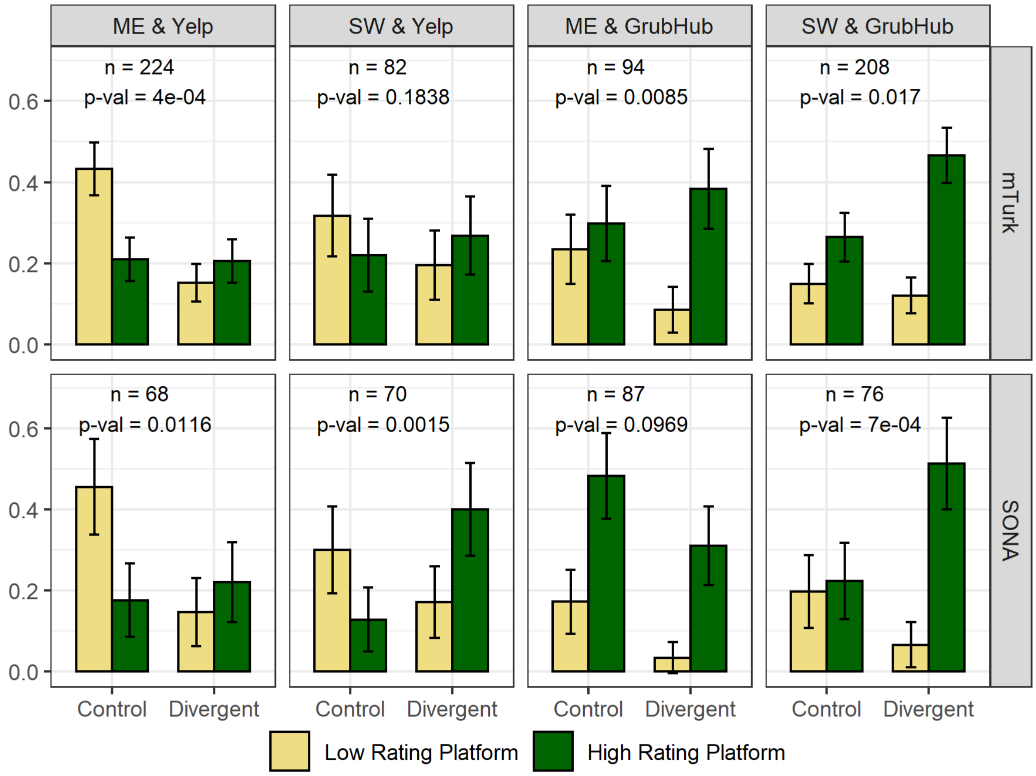


**Figure 1** **Subject choices in stated preferences studies 1 and 3**

for the corresponding platform–restaurant combination shown in brackets. 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.

**Table 2** **Subject choices in incentive aligned studies 2 and 4**

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

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 $10^6$ 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 $\approx 1$) treatment.[3] 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 the presence of a vertical spillover effect.

### 5.3. 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 US-based mTurk 608 workers who receive a $2 reward for participating in our study.

First, we replicate our results about platform choice in 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 remarkably 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 remarkably similar and 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

[3] Fisher exact test that is also commonly used with small sample sizes (Hyndman and Embrey, 2017) gives very similar results: *YelpHigh* p-value is 0.0210, *YelpHigh2* p-value is 0.0539, and *GrubHubHigh* p-value $\approx 1$.

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 ratings. Diverging average ratings across platforms may thus be an artifact of variation is 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 where 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 faces 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 ($2 \times 3$) 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 3 shows the percentage of subjects who chose the restaurant

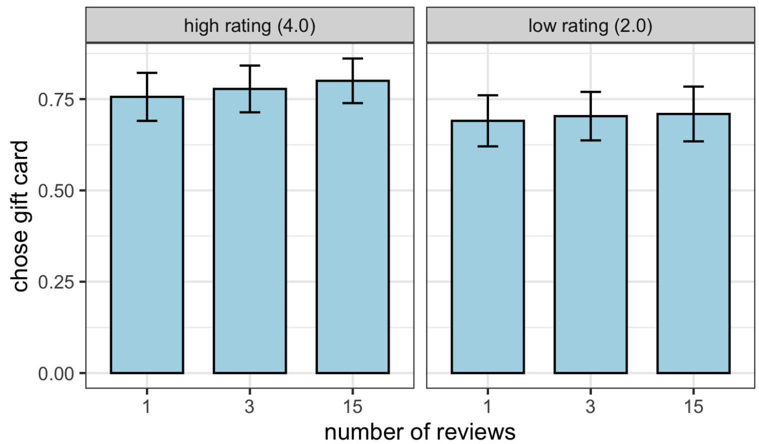


**Figure 2 The impact of the number of ratings on the choice between a gift card and a cash reward**

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.

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 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.

### 5.4. 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 2 shows the number of subjects who chose each option in Study 4, with the average rating for the corresponding platform–restaurant combination shown in brackets. 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 $10^6$ 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.

## 6. 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 mechanisms 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 pool showing that consumers are more likely to transact with a 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 an 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.

### 6.1. Managerial Implications

Our results show that a threat from hosting lower average ratings than competing platforms is real when both sides of the platform multihome, since consumers prefer to transact on a platform where the chosen vendor or item is rated higher. Consequently, platform owners need to manage their ratings systems strategically.In the following, we consider managerial and broader implications to practice from our results.

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 should therefore to 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 vendor attempts to artificially boost their ratings. For instance, a platform owner may want to follow a sample of vendors on their own and competitors' platforms. If the valence and volume of ratings and reviews in the sample evolves differently on different platforms, the platform owner may need to decide 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). However, adopting any alternative to the standard 5-star system is probably a feasible option 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 (REF), and certain online labor platforms (Kokkodis 2021), where average vendor ratings have become clearly compressed toward the maximum of five stars. While one cannot expect individual platforms to voluntarily put in place mechanisms to curb ratings inflation if that means that a platform is going to lose out in the competition, a regulator could, at least in principle, consider separating 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 as 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 businesses that enter 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 sufficient number of ratings is collected. For example, the restaurant booking platform OpenTable is currently using this approach.

This over-reaction to averages 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 compressed toward the maximum, then the entrant whose average vendor ratings will likely vary more is at particular disadvantage against more established competitors. Such problems could again be mitigated by using a third-party ratings provider.

### 6.2. 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 although not particularly surprising 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 due to the 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. Our experimental design controls for the platform attributes that are part of the treatment as nuisance variables, and for *a priori* preference for GrubHub over Yelp 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. 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 does not explain 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 offers a strategic incentive for platform owners to tolerate behavior that contributes to ratings inflation. The result 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 vendors who manipulate their ratings. This means that the average ratings are particularly prone to become inflated whenever both vendors and consumers multihome on several platforms, eroding the usefulness of the standard 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 customer surveys often provide input for important marketing and design decisions. By contrast, there are a number of 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.

### 6.3. Concluding Remarks

In this study, 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 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`.

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

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

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 JG (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.

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

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

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

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

Cho HS, Sosa ME, 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 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.

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

Filippas A, Horton JJ, Golden J (2020) Reputation Inflation.

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

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.

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.

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 YJ, Hosanagar K, 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):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.

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

Lin JJ, Chang CH, 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):438–458.

McGee P (2020) Apple: how app developers manipulate your mood to boost ranking. *Financial Times* URL `https://www.ft.com/content/217290b2-6ae5-47f5-b1ac-89c6ccebab41`.

Moe WW, Schweidel DA (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, 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, Xie J (2021) The fateful first consumer review. *Marketing Science* 40(3):481–507.

Parker G, Van Alstyne M, Jiang X (2017) Platform ecosystems: how developers invert the firm. *MIS Quarterly* 41(1):255–266.

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

Roehm ML, Tybout AM (2006) When will a brand scandal spill over, and how should competitors respond? *Journal 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.

Saifee DH, Zheng Z, Bardhan IR, Lahiri A (2020) Are online reviews of physicians reliable indicators of clinical outcomes? A focus on chronic disease management. *Information Systems Research* 31(4):1282–1300.

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.

Schneider C, Weinmann M, Mohr PNC, vom Brocke J (2021) When the stars shine too bright: the influence of multidimensional ratings on online consumer ratings. *Management Science* 67(6):3871–3898.

Siering M, Janze C (2019) Information processing on online review platforms. *Journal of Management Information Systems* 36(4):1347–1377.

Sun M (2012) How does the variance of product ratings matter? *Management Science* 58(4):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 MM, Cavusoglu H, Raghunathan S (????) Online product reviews: is a finer-grained rating scheme superior to a coarser one? *MIS Quarterly* .

Turner E, 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, Lambrecht A (2021) The effect of individual online reviews on purchase likelihood. *Marketing Science* 40(4):708–730.

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.

Wu J, Zhao H, Chen H (2021) Coupons or free shipping? Effects of price promotion strategies on online review ratings. *Information Systems Research* 32(2):633–652.

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.

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, 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.

# Appendix. Data Collection Instrument

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<br>**Average rating:** 4.5 stars<br>**Delivery hours:** 10:00am–11:45pm<br>*Delivery is fulfilled by Grubhub* | **Category:** Dinner, Lunch, Middle Eastern, Pitas, Smoothies and Juices<br>**Average rating:** 3.5 stars<br>**Delivery hours:** 10:00am–12:00am |
| **Restaurant 2** | **Category:**Sandwiches, Cheesesteaks<br>**Average rating:** 4 stars<br>**Delivery hours:** 11:00am–10:15pm<br>*Delivery is fulfilled by Grubhub* | **Category:** Cheesesteaks, Dinner, Hot Dogs, Lunch Specials, Sandwiches, Subs<br>**Average rating:** 4 stars<br>**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

**Figure 3** **Subjects' choice screen in the incentivized experiments.**