ECONSTOR
Make Your Publications Visible.

A Service of
ZBW
Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

Cicognani, Simona; Figini, Paolo; Magnani, Marco

Working Paper

Social Influence Bias in Online Ratings: A Field Experiment

Quaderni - Working Paper DSE, No. 1060

Provided in Cooperation with:
University of Bologna, Department of Economics

Suggested Citation: Cicognani, Simona; Figini, Paolo; Magnani, Marco (2016) : Social Influence Bias in Online Ratings: A Field Experiment, Quaderni - Working Paper DSE, No. 1060, Alma Mater Studiorum - Università di Bologna, Dipartimento di Scienze Economiche (DSE), Bologna, https://doi.org/10.6092/unibo/amsacta/4669

This Version is available at:
https://hdl.handle.net/10419/159898

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

https://creativecommons.org/licenses/by/3.0/

WWW.ECONSTOR.EU

Mitglied der Leibniz-Gemeinschaft

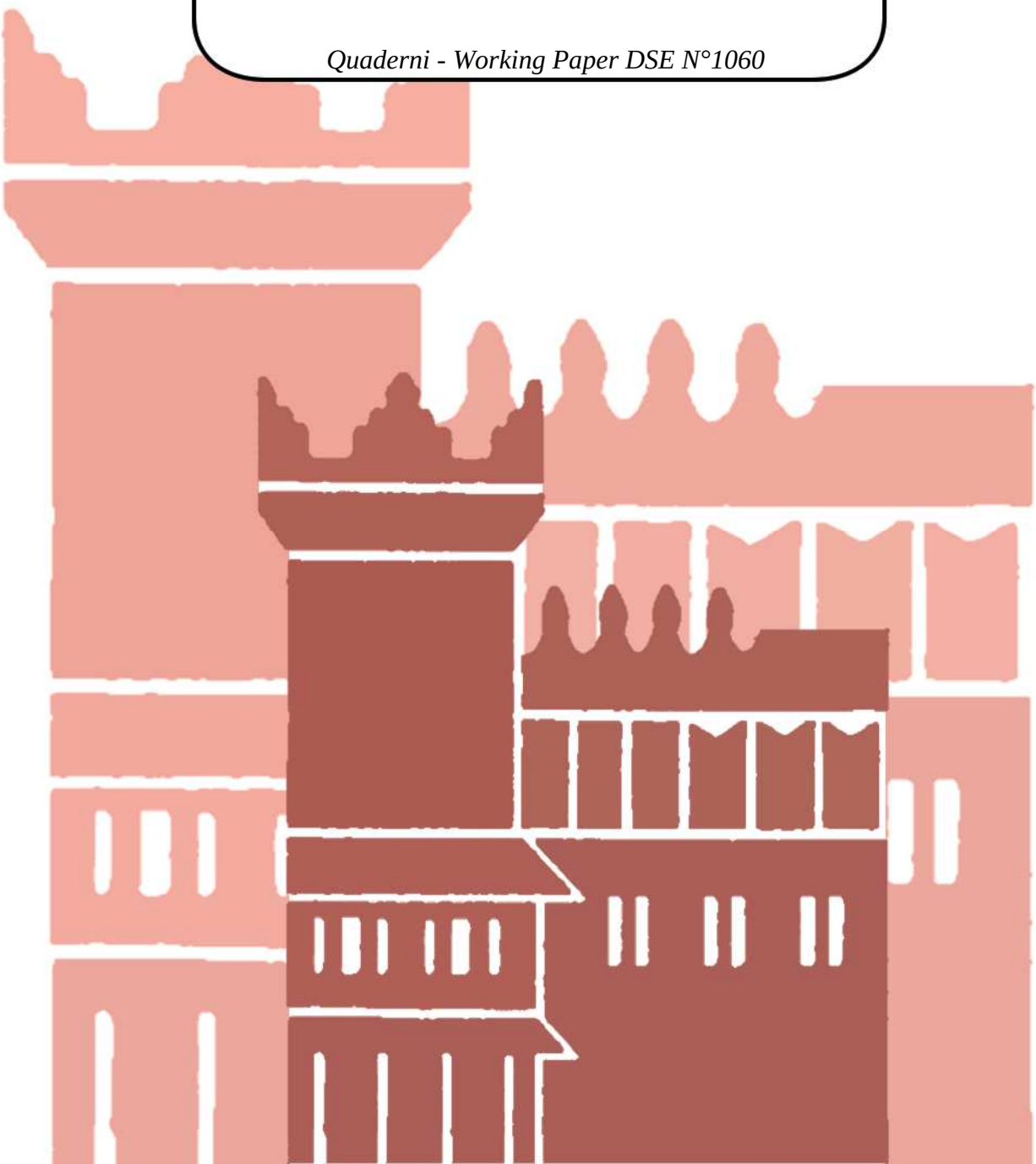
ISSN 2282-6483

Alma Mater Studiorum - Università di Bologna
DEPARTMENT OF ECONOMICS

Social Influence Bias in Online Ratings: A Field Experiment

Simona Cicognani
Paolo Figini
Marco Magnani

Quaderni - Working Paper DSE N°1060



Social Influence Bias in Online Ratings: A Field Experiment

Simona Cicognani* Paolo Figini† Marco Magnani‡

February 18, 2016

Abstract

The aim of this paper is to study the empirical phenomenon of rating bubbles, i.e. clustering on extremely positive values in e-commerce platforms and rating web sites. By means of a field experiment that exogenously manipulates prior ratings for a hotel in an important Italian tourism destination, we investigate whether consumers are influenced by prior ratings when evaluating their stay (i.e., social influence bias). Results show that positive social influence exists, and that herd behavior is asymmetric: information on prior positive ratings has a stronger influence on consumers' rating attitude than information on prior mediocre ratings. Furthermore, we are able to exclude any brag-or-moan effect: the behavior of frequent reviewers, on average, is not statistically different from the behavior of consumers who have never posted ratings online. Yet, non-reviewers exhibit a higher influence to excellent prior ratings, thus lending support to the social influence bias interpretation. Finally, also repeat customers are affected by prior ratings, although to a lesser extent with respect to new customers.

Keywords. Online Ratings, User-Generated Contents, Field Experiment, Rating Bubbles, Social Influence Bias, Herd Behavior
JEL codes: C93, D83, L86, Z31.

Acknowledgements: The authors thank participants at the 2015 Consumer Behaviour Tourism Symposium held in Munich and participants to the seminars held at University of Bologna and University of Lugano. We gratefully acknowledge the research assistance of students of the Master in Tourism Economics and Management, University of Bologna, Rimini Campus and of Riccardo Tonielli. The usual disclaimers apply.

*Corresponding author, University of Verona, simona.cicognani@univr.it
†University of Bologna, paolo.figini@unibo.it
‡University of Bologna, marco.magnani18@unibo.it

1 Introduction

Nowadays user-generated ratings are an inseparable part of the Web - an essential component of what was called the Web 2.0 by many. They drive and complement customers' behavior and experience in many diverse economic sectors, from tourism services like accommodation and restaurants to physical products such as electronic devices and printed books; from digital goods such as e-books, videos and songs, to political opinions and personal preferences. Everything can be rated everywhere on the Internet: through review platforms (e.g. Yelp, TripAdvisor) - businesses mainly devoted to collect and aggregate user-generated contents (hereafter UGCs); and through rating systems developed within e-commerce platforms (e.g. Amazon, eBay, Booking.com), databases (e.g. IMDB, Glassdoor), and social media (e.g. Facebook, Twitter, Disqus). Rating systems allow people to express and publish their opinion through either a binary scale (like / dislike) or, more commonly, Likert-type scales (the most popular representing a satisfaction level ranging from 1 to 5 points); and they often include the possibility of writing a small review. Rating systems had become so relevant that companies started to enrich them by including pictures uploading, several sub-categories which can be rated as well, surveys, the possibility of sharing and even rating reviews (e.g. Amazon "Was this review useful?") and reviewers themselves (e.g. eBay feedback system).

The absolute importance of user-generated online ratings for businesses and social relationships in daily life drives the motivation of this paper. Together with being a relevant issue for data science, the ubiquity of online ratings has important implications in management and marketing, in the organization of markets and in the behavioral pattern of users-customers. According to a 2015 Nielsen report[1], consumers' online opinions are trusted by 66% of customers - the third most-trusted source of information on products and services after recommendations from friends or relatives, and branded websites. In general, consumers tend to trust this type of word-of-mouth information because of the lack of commercial self-interest, which might bias information coming from other sources, such as intermediaries or companies (Litvin et al., 2008). The relevance of online ratings has recently propelled a thorough investigation about their integration within the official classification of services (see the UNWTO Annual Report 2014 in the hotel sector[2]). Hence, it is an important topic worth investigating in socio-economic research.

However, online ratings are not trouble-free and call for specific investigation. So far, the recent literature has focused on four main issues related to UGCs: i) the drivers of consumer-to-consumer (C2C) online activity and of individual rating behavior; ii) the analysis of social influence when reviewing online, in order to estimate how much users are affected by prior ratings and other information already available in the web site; iii) the properties of the aggregate distribution of ratings: if there are biases in how users review products and services, online ratings

[1] "Nielsen: Global Trust in Advertising" (September 28, 2015); available at: http://www.nielsen.com/us/en/insights/reports/2015/global-trust-in-advertising-2015.html

[2] "World Tourism Organization UNWTO: Annual Report 2014" (2015); available at http://www2.unwto.org/annualreport2014

cannot deliver efficient and trustworthy information; iv) the impact of online ratings on behavioral intentions and product sales. Each of these issues has positive but also normative contents, suggesting reforms for: improving the reliability of user-generated online ratings; the organization and functioning of online rating platforms; enhancing the effectiveness of firms' marketing policy; improving the regulation of markets in order to enhance efficiency.

This paper focuses on the second issue: investigating whether, and how, individual rating behavior is affected by prior aggregate ratings. The interaction between individual behavior and prior rating information might be one of the drivers for the empirical phenomenon of rating bubbles, i.e. the clustering of ratings on extremely positive values within e-commerce and rating websites. Several reasons have been proposed by the literature to explain this stylized fact: purchasing bias, under-reporting bias, observational learning, herding behavior. We conducted a field experiment in order to identify and isolate the impact of social influence bias (hereafter SIB) on users' ratings. The experiment was conducted during Summer 2015 in Rimini, a key Italian destination for tourism, ins one of the most important sector for user-generated online ratings: accomodation. The experiment consisted in handing out questionnaires to hotel customers after their stay, asking them to rate the hotel on a 5-point scale. The main novelty of this research is twofold: on the one hand, we focused on a typical experience good, where subjects self-selected themselves into the market; this is in direct contrast with most of previous studies where individual rating behavior was analyzed for search goods, non-market items such as political opinions or personal comments;[3] or through lab experiments where the product was given to subjects by the experimenter. On the other hand, the experimental design allowed us to assess any different pattern in the behavior between frequent reviewers (posters) versus non-reviewers (lurkers - a terminology accepted in the literature).

The paper is structured as follows: Section 2 provides a literature review on online rating systems; Section 3 motivates the paper and unfolds the main research questions and the novelty of our approach; Section 4 describes the experimental design; Section 5 presents the main results; Section 6 discusses the findings and relates them to previous and future research.

2 Literature review

2.1 A General Assessment

The literature about user-generated online ratings has recently been flourishing, ranging in terms of topics, sectors under investigation and methodology. Research papers mainly clustered within four main questions: i) which factors affect people's rating behavior online? ii) Is it possible to disentagle genuine personal tastes and experiences

[3] Ong et al. (2015) showed how the sense of ownership plays a relevant role when consumers are evaluating a product; therefore, it is important to discriminate between products that are purchased and consumed and those that are not.

from the community norm in the final rating? iii) What are the properties of rating distributions and how much can they be trusted? iv) What is the impact of online ratings on product adoption and sales?

Regardless of the topic, these research studies were conducted for ratings in a wide array of contexts: from political opinions (Krishnan et al., 2014) to entertainment products like movies (Schlosser, 2005; Lee et al., 2015) and music (Hu et al., 2009); from software (Duan et al., 2009) to physical products like printed books (Godes and Mayzlin, 2004), to services like restaurants (Cai et al., 2009) and accommodation (Yacouel and Fleischer, 2012). As regards the methodology, parallel to the use of laboratory experiments (Schlosser, 2005; Hu et al., 2009) and field experiments (Cai et al., 2009), population-scale networked datasets recently enabled novel investigations on the diffusion of information through in-vivo experiments on rating platforms within social networks (Aral and Walker, 2012; Muchnik et al., 2013). Empirical data have also been used to study rating behavior patterns on the Internet (Krishnan et al., 2014). In the remaining of this section, we review the literature by research area, recalling the differences in sectors and methodologies when appropriate.

2.2 The Drivers of Individual Rating

This review does not aim at investigating why people decide to write and post reviews online - the motives driving individual behavior are only tangential to the research question of this paper.[4] As any other voluntary activity, reasons are somehow connected to intrinsic motivations, self-signaling, reputation-building, willingness to reciprocate, altruism. But there could also be underlying strategic reasons operating, due to the willingness to push towards the improvement in service/ product quality by posting a bad online review. Hence, the analysis of the difference between posters and lurkers in terms of motivation and incentives is another topic the literature is currently tackling (Lai and Chen, 2014; Liao and Chou, 2012). Differently, we consider the strand of literature investigating the content of reviews and those factors affecting ratings.

Given the heterogeneity of users, tastes and experiences, the ideal situation would have ratings reflecting the true and subjective individual's opinion. Since posters do not live in isolation but interact and communicate in social networks, and are exposed to a myriad of signals carrying relevant information, it is of fundamental importance to distinguish the impact on rating behavior coming from genuine individual experience (if anything like this can be identified) from that stemming from the adoption of informational or behavioral patterns within a network. In particular, people interacting in the same network often share similar views, and have correlated preferences because of: i) the reflection problem (Manski, 1993; Bramoullé et al., 2009), which is related to homophily, i.e. the tendency of individuals to socialize in circles of people with similar tastes; ii) simultaneity, the tendency of connected individuals to co-influence each other and behave similarly at the same time; iii) confounding factors,

[4] See **?** for a detailed theoretical study with a laboratory validation on this issue.

which lead to similar responses to the same exogenous shocks (for example marketing strategies implemented by firms or changes in product characteristics). Yet, while these factors imply correlation about ratings but not bias, it is important to isolate and measure a fourth factor: iv) peer influence, which is related to social and observational learning in the network and might imply contagion in the building of the aggregate opinion. The latter may lead to informational cascades and herd behaviors that might bias the aggregate score, with interesting implications to analyze. [5]

2.3 Social Influence Bias in Online Ratings

Since Asch experiments (Asch, 1951, 1955), social influence has been extensively studied and debated. In this line of research, it is fundamental to establish whether individuals' decisions are affected by the pure observation of others' actions (observational learning) or by the informational content observed in others' actions (social learning). The susceptibility of recommendation systems to rating biases, as it has been reassessed by Cosley et al. (2003), points out that the true perception about a rated item can be biased. If individuals are highly susceptible, early adopters of a product or influential opinion leaders are pivotal in the diffusion of information and behaviors, thus possibly generating a SIB: the tendency to conform to the perceived norm in the community.

When applied to online reviews, the informational cascade stemming from prior ratings or previous comments might generate a significant bias in subsequent individual rating, thus creating herd behavior, a phenomenon in which individuals converge to the same outcome. Since herd behavior leads to sub-optimal market equilibria (Banerjee, 1992, 1993; Bikhchandani et al., 1992, 1998), informational cascades and bias in the perception of quality might imply that "aggregate collective judgement and socialized choice could be easily manipulated, with dramatic consequences for our markets, our politics, our health" (Muchnik et al., 2013, p.647).

Despite the large theoretical literature on observational and social learning, the empirical evaluation of these effects is not an easy task, albeit highly relevant in terms of efficiency and welfare implications. Empirically, herd behavior has been observed in many instances (Anderson and Holt, 1997; Çelen and Kariv, 2004) but the main challenge comes from the co-existence of the aforementioned confounding channels (homophily, simultaneity and other confounding factors). In general, herding is a possible outcome even when it contradicts one's own private signals due to: i) rational expectation of making fewer mistakes when following the crowd; ii) lower mental effort involved in the decision; iii) fear of loss of reputation when dissenting from the majority. Herding stems from the relative importance attributed to others' prior information compared to own information and experience. With

[5] Although informational cascades and herd behavior are often used as synonymous, there is a slight difference in their concept (Çelen and Kariv, 2004): an informational cascade occurs when a sequence of individuals ignore their private information when making a decision. Herd behavior arises when a sequence of individuals make the same decision, not necessarily ignoring own private information. Hence, an informational cascade implies herding, but herding is not necessarily implied by an informational cascade.

herding, in a pre-valence setting, users would choose a product or assume a behavior because of the informative content of observed others' actions; in a post-valence setting, users would adjust their ratings downward after observing prior negative reviews, in a sort of self-presentational concern (perhaps because negative reviewers are often perceived as more competent and intelligent, as explained by Schlosser (2005)).

In many cases, individuals may be led to mimic others' (observed) choices simply because of conformity concerns (Cai et al., 2009).[6] Conformity has been found to be associated with: i) the issue under discussion: the fear of dissenting becomes more salient when opinions are expressed, less salient in product adoption; ii) a popularity issue: bandwagon effects become prominent for popular products, less important for niche products, for which people are more willing to differentiate themselves (Dellarocas et al., 2010; Berger and Heath, 2008; Hu and Li, 2011; Lee et al., 2015). Along this argument, Moe and Trusov (2011) found that the posting of early positive ratings encourages negative ratings in a sort of differentiation and correction effect. They also found that when ratings start being polarized, subsequent raters tend to be more moderate (to avoid supporting one position and going against the opposite one), something that Schlosser (2005) called "multiple-audience effect". This in turn leads to a steady-state where the number of reviews is negatively associated with its average score, also because of the growth in reviewers' heterogeneity (Godes and Silva, 2012; Lee et al., 2015). Finally, observing the choices made by others also makes these choices important. It is hence possible that individuals follow others' choices simply because these are more salient than alternative choices. Many other papers found evidence of herding (Wu and Huberman, 2008), and imitation is particularly relevant for sporadic reviewers who more often imitate previous reviewers in a sort of bandwagon effect (Moe and Schweidel, 2012).

In this line of research, the recent use of in-vivo randomized experiments has become popular. Using this methodology, Muchnik et al. (2013) found asymmetric herding: whereas positive social influence accumulates, creating a tendency toward rating bubbles (the likelihood of positive ratings increased by 32% as a consequence of a manipulated initial positive rating, leading to an increase in the final rating by 25% on average), negative social influence inspired users to correct manipulated negative ratings. Krishnan et al. (2014) studied individual behavior in a rating system evaluating California state policies (the California Report Card), where the individual rating builds in three phases: in the first one, users rate the policy; then they observe the median rating; finally they are allowed to change their own evaluation (although they are not explicitly told to be allowed to do so) before submitting the final score. Social influence bias is found to be relevant, since 35% of users changed their rating after the observation of the median and, for this group, final ratings were significantly more concentrated around the median than initial ratings, the effect being symmetric for those with initial ratings above and below the median.

[6] In a laboratory experiment, Cicognani and Mittone (2014) investigate whether imitation is higher for tasks that entail higher cognitive costs. Interestingly, subjects do not conceive imitation as cognitive short cut, and imitative behaviors are driven by subjects who are under-confident about their own skills.

2.4 The Aggregate Distribution in Ratings

These carefully designed experiments demonstrate that being exposed to prior collective opinions distorts both individuals' decision making as well as own perceptions of quality and value, in turn generating informational cascades that might be irrational and pervasive (Wang et al., 2014). If there are biases in how users rate products and services, online ratings cannot deliver efficient and trusted information, a problem for both users and websites. One of the most striking empirical regularities that has been observed within online rating systems is the so-called J-shaped distribution: the clustering of ratings around extreme values, following a bimodal distribution with a high frequency of extremely positive ratings and a lower frequency of extremely negative ratings. The terminology was created by Hu et al. (2009) by looking at Amazon rating distributions across several product categories; nevertheless, this phenomenon has been reported since the early 2000s on a wide array of rating platforms. As Hu et al. (2009) pointed out, in such cases the average rating might not reflect the true distribution of opinions in the population of interest - and therefore might not be an adequate proxy of product quality. This is fairly intuitive: if ratings are clustered towards extremely positive values, it will be harder to distinguish a high-quality product from a good product because their averages would not differ substantially.

The literature proposed several reasons to explain the J-shaped distribution phenomenon. For example, online ratings tend to be extremely positive because of purchase self-selection, while bimodality is due to under-reporting bias (Hu et al., 2009). In fact, the authors compared the actual Amazon rating distribution (clustered on extremely positive values) of a randomly selected music CD with the rating distribution of the same product obtained in a controlled laboratory setting, which turned out to be bell-shaped, i.e. unimodal. The authors explained this discrepancy by pointing out that actual buyers are more likely to be positively predisposed towards a product - they had voluntarily purchased it (self-selection bias); and that customers with extreme preferences are more likely to express their opinions than those with moderate views, who might not bother to write a review (under-reporting bias). Another factor explaining rating bubbles, as proposed by the literature, is, indeed, herd behavior, which can act as a reinforcing effect, or as a stand-alone driver. Towards this direction, Muchnik et al. (2013) showed that SIB is quite well present in social networks where product rating is implemented. Finally, fraud might be another reason behind this phenomenon since fake ratings are typically concentrated at the top and bottom of the distribution (Luca and Zervas, 2013); in fact, fraud distortions might also work together with SIB in order to create and reinforce J-shaped distributions; for example, a few initial and extremely positive fake reviews might trigger herding behavior and finally create a biased rating distribution. One can also conclude that, if SIB is really present, a cheater needs only to post a few fake reviews, and let the herd behavior do the rest.

One of the main recent developments of rating systems is social filtering; that is, to show or highlight reviews

posted by individuals within the user's social networks circles. In this way, it is possible to investigate any differential impact of prior ratings by the crowd (i.e. the overall population of rating individuals) and by friends (as identified by the social network), overcoming the important issue of anonymity of reviews (Dellarocas, 2003). Muchnik et al. (2013) found that propagation of rating bubbles is quicker when positive comments are posted by friends. This, in turn, reinforces the effect played by homophily, clustering ratings on one mode. Similarly, Lee et al. (2015) studied an important movie platform and found evidence that friends' ratings always induced herding. They also highlighted that the existence of social networking reduced the likelihood of herding on crowd and of an "audience size" effect: ratings scores were positively associated to the number of friends.

2.5 The Impact of Rating Systems

The most important economic consequence of UGCs in online rating systems is on market equilibria. Four issues become relevant: customer behavior, product adoption, reputation and marketing.

Many studies analyzed how online reviews impact on behavioral intentions. Firstly, reading product recommendations is a strong sign of behavioral intentions (Senecal and Nantel, 2004): in the absence of recommendations, willing-to-purchase consumers conducted even more informational searches (Smith et al., 2005). Secondly, consumers were more influenced by recommendations when dealing with an experience product (e.g. wine) than with a search product (e.g. calculator) (Senecal and Nantel, 2004). Thirdly, Park et al. (2007) found that consumers' purchasing intentions increased with the number of reviews, since this suggests that the product is popular. By means of an online experiment, Viglia et al. (2014) demonstrated how the number of reviews becomes a valuable signal even if the average rating is low. Fourthly, Zhang et al. (2010) studied the behavioral intentions of Chinese customers after reading reviews posted on a Chinese rating platform, finding that perceived informativeness and argument strength were additional positive drivers of behavioral intentions, together with the volume of reviews (differently from source credibility, which was found not to be an important factor). Finally, disclosure of identity was another important factor of behavioral intentions (Liu and Park, 2015).

As regards product adoption, since reviews are an acknowledged proxy for quality, ratings are expected to directly influence the quantity of sold units. While in the early age of UGCs the novelty and the lack of trust were perceived as a problem and were not robustly linked to sales (Chen et al., 2004), recent literature found more consistent results. Many works found evidence of a positive correlation between review scores and sales (Chevalier and Mayzlin (2006); Li and Hitt (2008), in the book industry). Liu (2006) and Duan et al. (2008) found a positive correlation between the number of reviews, rather than their scores, and sales. Dellarocas et al. (2007) also found that the volume, together with the valence and the dispersion of online ratings, are all positively associated to future sales in the movie industry. Chintagunta et al. (2010) determined that sales are positively associated to reviews

when they come from own networks, while Forman et al. (2008) found that the association is stronger when the identity of the reviewer was disclosed. In contrast, Duan et al. (2009) did not find any link between user ratings and software adoption.

Using in-vivo randomized experiment, Aral and Walker (2012) controlled for unobservable factors such as latent homophily to identify and separate the effects played by influential members in a network and susceptibility in product adoption. They profiled susceptible and influential users exchanging information and opinions in a movie-related commercial Facebook application and showed that the joint distribution of influence, susceptibility and the likelihood of spontaneous adoption in the local network determined the diffusion of behaviors.

In a field experiment, Cai et al. (2009) showed that, in the context of restaurant dining, demand for the top five dishes was increased by 13-18% when they were revealed to the customers as being the most popular ones in the previous week. In contrast, being merely mentioned as some sample dishes did not significantly boost their adoption. Hence, the authors concluded that the effect is due to observational learning rather than saliency. Moreover and consistent with theoretical predictions, they found some modest evidence that observational learning effect was stronger among infrequent customers.

Ratings are not only important for product adoption and sales. Good scores from previous customers signal high quality and could lead companies to ask a premium price. Rating platforms and e-commerce systems are an important source of reputation particularly for services or products in which the assessment of quality is more difficult and volatile, like experience goods. Yacouel and Fleischer (2012) found that, holding everything else constant, high rating scores are associated with higher prices for the hotel sector in an important e-commerce platform, Booking.com. Similarly, de Albornoz et al. (2011) reported that user-generated ratings have a significant impact on purchasing decisions, so that consumers are willing to pay about 20% more for services receiving the highest score than for similar services receiving a slightly lower score.

Given the premium on prices and sales stemming from good reviews, online rating platforms provide strong incentives to fraud. Luca and Zervas (2013) investigated the economic incentives to commit review fraud on Yelp: roughly 16% of reviews on the web site are considered fake by the internal filtering algorithm. These reviews tend to be more extreme (positive or negative) than "genuine" reviews, and the share of suspicious reviews has grown significantly over time. Moreover, suspicious reviews are negatively associated with the number of reviews and positively associated with the number of recent bad reviews. Overall, Luca and Zervas (2013) suggested that suspicious reviews with a positive content are associated with negative changes in the company's reputation, while suspicious reviews with a negative content are driven by stiff competition in the sector.

2.6 Recommendations

The literature also provides many recommendations to mitigate or eliminate the problems encountered in online rating systems and UGC platforms; these management implications would enhance the effectiveness of marketing strategies and better regulate the market. The main question is how to reduce or eliminate any kind of distortion in online ratings in order to unfold the true quality of goods. While Aral (2014) proposes systems where the average score is hidden while users rate a product (quite an unrealistic solution), Krishnan et al. (2014) suggest to introduce the 3-step rating system applied by their case-study, the California Report Card, and then use machine learning to estimate the social influence bias and to correct it before the review is posted onto the platform. Differently, Wang et al. (2014) develop an algorithm aimed at predicting the dynamics of review scores, thus helping to measure and isolate the impact of social influence and herding. Finally, systems that disclose and check the identity of reviewers mitigate the incentive to cheat, while the connection with individual social networks allows users to filter those ratings coming from own circle of acquaintances, thus increasing trust and homophily through the generation of "personalized"aggregate ratings.

3 Contribution and Hypotheses

In this *mare magnum* of empirical and experimental evidence, this work attempts to shed light on some mixed results and under-investigated issues. First, the assessment of SIB in individual rating behavior: the literature still has to find solid results which can confirm herd behavior as the primary driver of biased rating distributions. Second, most of the empirical and experimental literature mainly focused on circumstances that are not relevant when it comes to understand consumers' behavior. In fact, some authors analyzed SIB in opinion rating platforms (users' comments to news as in Muchnik et al. (2013); for political issues as in Krishnan et al. (2014)), where users are not required to purchase the object they rate nor they actually consume it. Lee et al. (2015) considered online ratings in a social movie web site using field data - without having the possibility to verify whether movies were actually purchased or even watched. On the other hand, laboratory experiments focused on the rating phase without taking into account the issue of self-selection into the market: both Schlosser (2005) and Hu et al. (2009) used controlled experiments to study individual rating behavior by assigning the product to subjects. Therefore, subjects had to rate a product that they had not chosen (and paid for), making the outcome less salient - this shows in the main results, as illustrated in Subsection 2.4.

To the best of our knowledge, this is the first experiment studying individual rating behavior and SIB in the relevant case of an experience good, and isolating the problem of self-selection by investigating only individuals who actually consumed the good. In the case of experience goods, people find difficult to assess the quality before

consumption (differently from standardized products - search goods - and from opinions); hence prior information about quality is highly relevant for purchasing choices. Consumers tend to rely on different forms of word-of-mouth (such as online UGCs) to obtain sufficient information and reduce their level of perceived uncertainty (Liu and Park, 2015). Tourism has been one of the first sectors developing online rating platforms and UGC systems, and tourism services (in particularly accomodation) are typical examples of experience goods. Therefore, focusing on such sector seems essential when studying rating systems and their inefficiencies.[7]

As regards self-selection, we conducted a field experiment in order to specifically study the rating behavior of individuals after their real experience. Our approach is similar to Cai et al. (2009) in terms of methodology and sector of investigation, but while they look at product adoption, our focus is on individual rating behavior. Then, we first want to assess if:

Hypothesis 1. *Consumers' rating behavior is affected by prior ratings: SIB is present in online rating platforms.*

Most previous research, as recalled in Section 2, found an asymmetric SIB, though with discordant results. Hence, we want to check whether:

Hypothesis 2. *SIB is asymmetric: information about positive prior ratings has a different effect with respect to information about negative or average prior ratings.*

In tourism, a relevant dimension through which the customer base is segmented concerns repeating purchasing behavior: repeat visitors (i.e. customers who repeat the experience in the same destination, hotel, or other leisure services) against non-repeat visitors. We argue that for repeat customers, hotel accommodation can be considered as a sort of search good, while new customers treat it as an experience good. The former have a more solid private information regarding the quality of the accommodation services than the latter. In particular, since they opted for repeated stays at the hotel, their evaluation of the hotel can be considered as quite established and sound. This leads us to hypothesize that:

Hypothesis 3. *Repeat customers are less likely to be influenced by prior ratings than new customers.*

Finally, despite the great availability of online and networked population-based datasets, our study differentiates from most of previous ones since it is a pseudo-online field experiment: the experiment itself was conducted online but subjects were contacted offline. In this way we were also able to include customers who are not used to read online reviews, not to mention to actively contribute to them. If observational learning drives customers to familiarize with online review systems and to recognize their limitations and possible biases, the rating behavior

[7] As regards tourism and hospitality, several studies have investigated the role of online reviews in the decision-making process and in product sales for general trips (Xiang and Gretzel, 2010; Vermeulen and Seegers, 2009), hotels (Sparks and Browning, 2011), and restaurants (Racherla and Friske, 2012). Luca (2011) showed how consumer reviews are less influential for chain restaurants, which already have firmly established reputations built by extensive marketing and branding, than independent restaurants.

of frequent customers would be different from the one of infrequent reviewers. Hence, following the suggested factors driving J-shaped distributions by Hu et al. (2009), the under-reporting bias would confound with SIB. Our setting allows us to distinguish SIB from under-reporting bias due to the inactivity of moderate customers. Hence, we expect that:

Hypothesis 4. *SIB is not confounded by under-reporting bias: the rating distribution of non-reviewers is not statistically different from the distribution of frequent reviewers, and both groups are influenced by prior ratings.*

4 Experimental Design

The goal of the field experiment is to investigate the consequences of being exposed to different information sets (including data about prior ratings for the same service) when rating an experience good. This section illustrates the experimental design and is structured in two parts. We first present the pilot study that preceded and framed the field experiment; this initial phase was helpful in understanding the critical parts of the design and to fine-tune treatments, questions and general accessibility to the questionnaire. Afterwards, we focus on the experiment itself, by explaining its design and treatments in greater details.

4.1 Pilot Study

The pilot started June 17, 2015 and ran until July 2, 2015. It consisted of a paper-based questionnaire administered to restaurant customers just after they had dined and paid. The restaurant is placed on the seafront of Rimini, a popular seaside destination for tourism in Northeast Italy.[8] The questionnaire contained questions related to the customer's overall experience at the restaurant, to be reported on a 5-point scale, as well as four sub-dimensions of the overall experience (food, value, service and atmosphere).[9] In addition, there were socio-demographic questions, questions regarding the customer's attitude towards online rating platforms, and prior experience in the same restaurant. The experimenters asked customers whether they would want to participate in a research about online rating systems, carried out by the University of Bologna; those customers who had accepted received the questionnaire; once filled in, they had to put it in an envelope, in order to avoid any concern about anonymity (no personal information such as name or date of birth was asked). It is worth mentioning that experimenters distanced themselves while subjects were filling in the questionnaire, and that the experiment was conducted outside the restaurant, such that the staff could not interrupt or influence the on-going experiment.

The pilot study consisted of two stages. In the first stage (June 17 - June 21) the Control treatment was imple-

[8]The restaurant TripAdvisor rating average is 4 out of 5, based on 282 reviews.

[9]The 5-point scale aimed at mimicking the rating scale used by TripAdvisor and other popular rating platforms.

mented: customers were asked to fill in a questionnaire which did not include any information about prior ratings of the restaurant. The purpose of the control treatment was to collect ratings in order to create treatments implemented in the second phase (June 23- July 2), that is, questionnaires with different information sets about prior ratings.

In the Control treatment, 59 observations were collected. Among these, 24 reported an overall rating of 5 points, 29 reported 4 points and 6 assigned an overall rating of 3 points to the restaurant. Following this information, we run the 4-point and 5-point treatments. In the 4-point treatment, the questionnaire was the same as in the Control, except for a sentence above the overall rating scale informing subjects that 29 previous customers gave a 4/5 to the restaurant.[10] Likewise, in the 5-point treatment, before expressing their overall rating to the restaurant, subjects could read that 24 previous customers gave a 5/5 to the overall experience.

All treatments followed a between-subjects design. 4-point and 5-point treatments were randomly assigned one after the other to customers, regardless the numerosity of the participating group. The response rate was about 40%, and it was based on the ratio between the number of customers who accepted to fill in the questionnaire and the overall amount of customers approached by experimenters at the exit of the restaurant.

Results from the pilot study were promising, albeit not entirely significant in statistical terms. As expected, the overall average rating in the Control treatment (4.36/5) was in between that of the 4-point and the 5-point treatments (4.19/5 and 4.45/5, respectively). Yet, only the difference in the overall average rating between 4-point and 5-point was statistically significant, albeit at a 10% level ($p = 0.090$, two-sample Wilcoxon rank-sum test, $N1 = 49$ and $N2 = 37$, two-sided).

Some issues arose during the pilot. The low variability in ratings may have been due to an experimenter effect: although distant from the customers when the questionnaire was filled in, the presence of the experimenters may have driven ratings upwards. Moreover, in some cases the questionnaires were incomplete. This may have been due to some tiring effect (as it was administered after dinner), non-compulsory answers, and the scarce lighting of the area outside the restaurant. Finally, for a restaurant already very well-rated on TripAdvisor, proposing not-much differentiated treatments could have been too weak in order to generate a substantial effect in the rating behavior. For these reasons, we decided to revise the experimental design, and conduct the experiment through a pseudo-online procedure, which enabled us to correct the three main issues encountered in the pilot: experimenter effect, compulsory answers and greater heterogeneity in terms of information sets across treatments.

[10] The exact sentence stated "29 among previous interviewed people gave a rate of 4 out of 5".

4.2 Field Experiment

The field experiment was conducted between August 3, 2015 and September 30, 2015 in a 3-star superior hotel located in Gatteo Mare, a seaside destination 20 km North of Rimini. It consisted of an online-based questionnaire (developed using Google Forms) handed out to the customers of the hotel after their stay through a flyer, containing a Web URL. The procedure of handing out the flyer was always conducted by the hotel manager, when customers were checking out; customers were informed they could fill in the questionnaire within two weeks using their own computers, smartphones or tablets.[11]

The questionnaire included three parts: in the first part customers were asked to rate the hotel (the overall experience, and four other categories: sleep quality, value, service, atmosphere); the second part included socio-demographic questions; the third part included questions about customer's experience in the same hotel and destination and own attitude towards online reviews (see Questionnaire in Appendix C, Figures 4-8). By completing the questionnaire, subjects were allowed to participate to a lottery (entry-free), whose prize consisted of a weekend stay at the same hotel in the following year.[12]

The experiment included three treatments, related to the set of information about prior ratings; the questionnaires were exactly the same, bar a sentence just above the overall rating question informing subjects about the rating attitude of previous customers. In the Control treatment, customers were asked to fill in the questionnaire without any information about prior ratings. In the 3-point treatment, subjects were made aware that at least 17 prior customers of the hotel expressed a 3/5 rating. Likewise, in the 5-point treatment, the information displayed before asking the overall rating was that at least 17 previous customers gave a 5/5 rating.[13] Figure 1 displays the first part of the questionnaire for the different treatments.

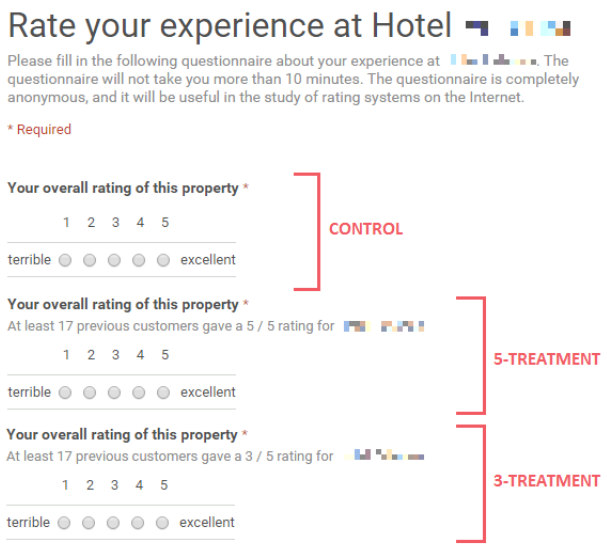
The three treatments were assigned randomly to the customers by the hotel manager: flyers containing the Web URLs related to the different treatments were handed out in sequence at the time of the check-out, independently on customers' characteristics. The potential selection bias is, therefore, orthogonal to treatments. Out of 400 flyers distributed, 75 questionnaires were completed: the response rate was hence 19%. After checking for missing or wrong codes, a total of 67 observations were considered in the analysis: 21 for the Control treatment, 22 and 24 for the 5-point and 3-point treatments, respectively.

[11] The flyer handed out to customers is reported in Appendix B, Figure 3.

[12] To avoid double-dippers, each flyer had a unique code that subjects had to report at the end of the questionnaire.

[13] At the time of the experiment, the hotel had an average rating of 4.5/5 across 232 reviews on TripAdvisor. 17 was the number of 3/5 ratings (there were, instead, 130 5/5 ratings), and was chosen to avoid deception.

Figure 1: Different informational sets across treatments



5 Empirical Analysis

This section illustrates the main results of the field experiment. We first present some descriptive statistics of the subject pool and the overall rating. We corroborate the analysis with non-parametric tests and several regression models considering as dependent variables both the rating scale itself and a dummy representing a 5-point rate.

The main characteristics of the customers taking part in the field experiment are displayed in Table 1 and, according to the hotel manager, they perfectly represent the hotel's clientèle.[14] The gender representation is balanced. The average customer is about 46 years old, Italian, with upper-secondary education. He mainly travels in couple or family, spending 7-10 days in the hotel. Customers are already quite acquainted with the destination and with the hotel: on average, they have already been five and four times to the same destination and to the hotel, respectively (8 customers had already been more than 10 times to the same hotel and destination). As for their rating attitudes, half of them has never written an online review, whereas one third has read a review about the hotel before booking the stay; 86% of them declared to have been affected by the reviews.

[14] A detailed description of the variables is provided in Appendix A, Table 9.

Table 1: Summary statistics - **selected variables**

Variable	**Mean** (Std. Dev.)	**Mode**
Female	0.48 (0.50)	
Age	45.58 (14.07)	
Italian citizenship	0.96 (0.21)	
Years of schooling	11.51 (2.84)	
Travel type		Family
Travel length		4-7 days
First time destination	0.27 (0.45)	
Repeat customer	0.27 (0.45)	
Type of stay		Half board All inclusive
Hotel advice		None
Not reviewer	0.52 (0.50)	
Review read	0.33 (0.47)	
Review source		TripAdvisor/ Yelp
Review influence	0.75 (0.88)	
Other prices	0.49 (0.50)	
Period of stay		10/08-23/08

Table 2: Summary statistics - **ratings**

Rating	**Mean** (Std. Dev.)
Overall	4.61 (0.70)
Sleep quality	4.49 (0.70)
Value	4.72 (0.60)
Service	4.81 (0.63)
Atmosphere	4.78 (0.65)

Does the subjects' rating behavior differ across treatments? A first clue is provided by Table 3. In line with Hypothesis 1, the overall rating mean is the lowest in the 3-point treatment and the highest in the 5-point treatment, with the mean of the Control treatment lying in between the two. Evidence reported in Table 3 suggests that being exposed to excellent prior ratings may have boosted subjects' overall rating. This is corroborated by Figure 2, in which the rating density under 5-point treatment is strikingly more skewed towards 5/5 with respect to 3-point and Control.

Table 3: Summary statistics - **overall rating** (by treatment)

Treatment	**Mean** (Std. Dev.)	**Observations**
Total	4.61 (0.70)	67
3-point	4.38 (0.82)	24
Control	4.52 (0.75)	21
5-point	4.95 (0.21)	22

This *prima facie* evidence is supported by non-parametric testing. When comparing the 5-point overall rating with the overall rating in Control, the difference is statistically significant at the 1% level (p = 0.007, two-sample

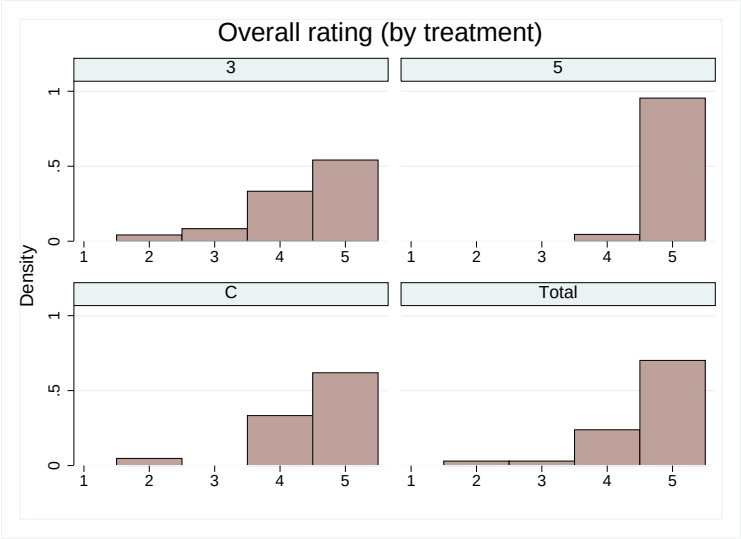
Wilcoxon rank-sum test, N1 = 22 and N2 = 21, two-sided). Similarly, also the difference between 3-point and 5-point is statistically significant at the 1% level (p = 0.002, two-sample Wilcoxon rank-sum test, N1 = 24 and N2 = 22, two-sided). This allows us to state:

Result 1. *Consumers' rating behavior is affected by information on prior ratings.*

We now turn to the type of influence of prior ratings (Hypothesis 2): is the social influence bias different in the case of exposure to positive prior ratings with respect to mediocre prior ratings? To answer this question, we compare the overall rating in 3-point *v.* Control, and 5-point *v.* Control. Only the latter is statistically significant (p = 0.007, two-sample Wilcoxon rank-sum test, two-sided), whereas the former is not (p = 0.520, two-sample Wilcoxon rank-sum test, two-sided).

Result 2. *SIB has an asymmetric effect towards positivity: being exposed to excellent prior ratings generates a significant positive bias in ratings, which is not mirrored by a negative bias after being exposed to negative prior ratings.*

Figure 2: Frequency distribution - **overall rating** (by treatment)



Hence, evidence of rating bubbles that is often observed stems from asymmetric herding. To corroborate these results, we estimated several models (Table 4) using as dependent variables: *Overall rating* (an ordered categorical variable taking values from 1 to 5) and a dummy variable called *Excellent rating* with value 1 if the overall rating is 5, and 0 otherwise. For each outcome variable, we estimated a non-linear model through ML (Model 1 is an ordered logit model, while Model 3 is a logit model) and a linear regression model through OLS (Model 2 and 4) as robustness checks.

Table 4: Treatment effects on rating behavior

	Ordered Logit (1)	OLS (2)	Logit (3)	OLS (4)
Dep. variable	Overall rating		Excellent rating	
(Intercept)	-	4.190***	0.116	0.529*
		(0.158)	(0.852)	(0.158)
5-point	2.958***	0.416***	2.648***	0.319**
	(0.521)	(0.042)	(0.114)	(0.036)
3-point	0.217	0.031	0.038	-0.010
	(0.390)	(0.088)	(0.093)	(0.023)
(Pseudo) R^2	0.225	0.313	0.243	0.257
Num. obs.	67	67	67	67

Notes: ***, **, * indicate the significance at the 1%, 5% and 10% respectively; standard errors are clustered at the treatment level.

All models in Table 4 include as control variables: gender; age; the stay period (considering three parts of the overall period: early August; mid-August, peak weeks of the tourist season; late August - September); being a new customer at the hotel; having read a review of the hotel; being an online reviewer. Goodness-of-fit measures in terms of McFadden R^2 in logit models and R^2 in linear models are fairly good, though not very large: 0.225 in Model 1; 0.313 in Model 2; 0.243 in Model 3; and 0.257 in Model 4. The F-test for the joint significance of regressors always leads to the rejection of the null hypothesis, which validates the models as specified.[15]

The significant coefficient of the 5-point treatment in the ordered logit model recalled in Table 4, column (1) shows the biasing effect of being exposed to information about excellent prior ratings;[16] the marginal effect is positive and equal to 0.42: by falling into the 5-point treatment, the probability of rating 5/5 increases by 0.42. On the contrary, being exposed to information about prior ratings below the true average does not create a statistically significant impact on the overall rating. It hence appears clear the asymmetric influence of positive and negative prior ratings on consumers' rating attitude. This is confirmed by Model 2 (Table 4, column (2)) where the 5-point coefficient is positive and statistically significant, while the 3-point coefficient is not.

The logit model recalled in Table 4, column (3) considers as dependent variable a dummy taking value 1 if the subject had given a 5/5 rating, and 0 otherwise. Here, the 5-point coefficient is positive and statistically significant, as shown by Table 4; the marginal effect is equal to 0.4, meaning that observing extremely positive prior ratings increases by a substantial amount the probability of giving a 5/5 rating to the property. The 3-point coefficient is not statistically significant, confirming the ineffectiveness of this treatment. As a further proof of this result, we also looked at the 5-point treatment odds ratio, that is the change in the ratio of probabilities of rating 5/5 when the treatment is in place and without; the odds of giving a 5/5 rating versus any other rating is 14.13 times greater,

[15] F-test results are discussed considering non-clustered standard errors; results are reported, instead, clustering standard errors at the treatment level.

[16] It is worth reminding that the 5-point treatment is a proxy for an extremely positive average rating on a UGC online platform such as TripAdvisor or Yelp. Likewise, the 3-point treatment is a signal of a mediocre average rating.

given that all of the other variables in the model are held constant. The odds ratio is statistically significant at the 1% level. Finally, we looked at the linear model estimated by OLS shown in Table 4, column (4); the 5-point coefficient is statistically significant and positive, corroborating all previous results.[17]

Does the individual rating attitude differ across different segments of the customer base, for example considering prior stays at the same hotel, or being used to online platforms already? As for repeat customers, a natural behavioral prediction is that they are less likely to be influenced by prior ratings than new customers (Hypothesis 3). We can provide an answer to this question by running an ordered logit model which includes two interaction terms[18]: between *Repeat customer*[19] and 5-point and 3-point dummies respectively. Along with a linear model estimated with OLS, results are shown in Table 5. Goodness-of-fit measures are not very large, but good nonetheless: McFadden R^2 in the ordered logit model is 0.280 while the standard R^2 in the OLS regression is 0.383.

Table 5: Treatment effects on rating behavior - repeat *v.* new customers

	Ordered Logit (1)	OLS (2)
Dep. variable	Overall rating	
(Intercept)	-	4.624^{***}
		(0.314)
Repeat customer	-0.931	0.051
	(0.680)	(0.197)
5-point	16.710^{***}	0.790^{***}
	(0.911)	(0.022)
5-point * Repeat customer	-13.637^{***}	-0.394^{***}
	(1.463)	(0.029)
3-point	-1.685^{***}	-0.378^{***}
	(0.408)	(0.044)
3-point * Repeat customer	3.472^{***}	0.723^{**}
	(1.151)	(0.182)
(Pseudo) R^2	0.280	0.383
Num. obs.	67	67

Notes: ***, **, * indicate the significance at the 1%, 5% and 10% respectively; standard errors are clustered at the treatment level.

Results show that being a repeat customer *per se* does not affect the rating behavior: the repeat customer coefficient is never statistically significant. On the other hand, though, the interaction terms are always statistically significant, for both treatments. This means that not only repeat customers are affected by prior ratings, but they are influenced differently with respect to new customers. In order to interpret ordered logit results (neither the sign or the magnitude can be inferred from estimated coefficients), and obtain a clearer picture of treatment effects on repeat customers, we computed predicted probabilities for the overall rating being 5/5 in each sub-segment of the

[17]Results are also robust to different specifications of the model, where control variables with insignificant coefficients are dropped from the regression.

[18]We did not conduct non-parametric testing because of the low numerosity of some sub-segments of the subject pool.

[19]This dummy variable takes value 1 if the subject has already been in the same hotel prior Summer 2015.

subject pool.

Table 6: Predicted Pr(*overall*=5) - repeat *v.* new customers (by treatment)

	5-point treatment	**3-point treatment / Control**
Repeat customer	0.954***	0.624***
	(0.006)	(0.019)
New customer	1***	0.566***
	(0.000)	(0.049)

Table 6 shows the predicted probability of subjects assigning 5/5 conditional on previous experience with the property. Since subjects' behaviors in the control group and 3-point treatment are not statistically different, the two groups have been incorporated as the alternative against the 5-point treatment. While the predicted probability increases in the 5-point treatment for both groups, repeat customers seem to be less affected by prior positive ratings. Indeed, the treatment effect is slightly less effective,[20] and repeat customers are already more likely to give an excellent rating.[21] This is intuitive: the opinion of a customer who chose again the same hotel must be already extremely positive. Yet, being a repeat customers does not allow subjects to escape rating biases.

Result 3. *Repeat customers are also affected by prior ratings.*

Result 4. *Repeat customers are less positively influenced by the 5-point treatment, and less negatively influenced by the 3-point treatment, than new customers.*

These results however must be handled with care: the repeat customer coefficient is not robust to different specifications of the model: by dropping the variables whose coefficients are not significant, the coefficient becomes positive and significant. The low number of observations is probably the reason behind these unstable estimates. Indeed, there are only 18 new customers in the subject pool, with 10 observations in the control group, 5 in the 3-point treatment, and only 3 in the 5-point treatment. This is also the reason why we could not estimate the logit model considering *Excellent rating* as dependent variable, and why we refrain from doing any non-parametric testing.

Thanks to the design of our experiment, the subject pool also included people who have never actively used online rating platforms: the lottery attracted customers who are not interested in user-generated online reviews *per se*, adding them to the cluster of people who on the contrary use websites such as TripAdvisor to rate goods. Understanding whether the rating attitudes of these two groups of subjects differ would help us shed light on one of the proposed explanations for rating bubbles: the "brag or moan" phenomenon, or under-reporting bias; that is, those who write online reviews have more extreme (and typically more positive) preferences than those who do

[20] $\Pr(overall = 5 \mid repeat = 1, treat_5 = 1) = 0.954 < Pr(overall = 5 \mid repeat = 0, treat_5 = 1) = 1.$

[21] $\Pr(overall = 5 \mid repeat = 1, treat_5 = 0) = 0.624 > Pr(overall = 5 \mid repeat = 0, treat_5 = 0) = 0.566.$

not bother in expressing and share their own opinion. If the behavior of the two types of subjects is not statistically different, then this type of bias would be excluded from the list of reasons that drives J-shaped rating distributions. Our sample is almost equally divided between non-reviewers (35 subjects, 52% of the sample) and reviewers (32 subjects, 48%). The average overall rating for non-reviewers is 4.59/5, versus 4.63/5 for reviewers.

We compared the rating behavior between the two types of subjects at the aggregate level, and within treatments (for example, testing whether the average ratings of reviewers and non-reviewers in the 5-point treatment are statistically different). Considering the average overall rating for the whole sample, the two groups are not statistically distinguishable (p = 0.557, two-sample Wilcoxon rank-sum tests, two-sided). By treatments, the difference in the overall rating of these types of customers is never significant (p = 0.897 for 3-point; p = 0.229 for 5-point; p = 1.000 for Control, two-sample Wilcoxon rank-sum tests, two-sided).

Table 7: Treatment effects on rating behavior - reviewer *v.* non-reviewer

	Ordered Logit (1)	OLS (2)
Dep. variable	Overall rating	
(Intercept)	-	4.333***
		(0.071)
Not reviewer	-1.693	−0.497*
	(1.166)	(0.138)
5-point	0.738	0.020
	(0.581)	(0.123)
5-point * Not reviewer	17.421***	0.713*
	(2.025)	(0.178)
3-point	−0.278*	−0.173*
	(0.143)	(0.054)
3-point * Not reviewer	1.044	0.386*
	(0.956)	(0.115)
(Pseudo) R^2	0.261	0.350
Num. obs.	67	67

Notes: ***, **, * indicate the significance at the 1%, 5% and 10% respectively; standard errors are clustered at the treatment level.

These results are corroborated by the regression analysis in Table 7, and predicted probabilities in Table 8, computed from ordered logit estimated coefficients. Having no experience in posting reviews online exerts, *per se*, a marginally significant effect on the rating behavior only in the OLS specification.[22] When interacted with the treatment variables, *Not reviewer* acquires significance in the 5-point interaction, while being not significant, or only marginally significant in the 3-point interaction term. Table 8 provides predicted probabilities for assigning an overall rating of 5/5 for reviewers *v.* non-reviewers. As in Table 6, predicted probabilities increase in the 5-point treatment for both types of customers. However, the difference in the treatment effect is much smaller for reviewers than for non-reviewers, suggesting that subjects used to posting online reviews may be less affected by previous

[22]The dummy variable *Not reviewer* takes value 1 if the subject has never rated a product online.

excellent information about the hotel.

Table 8: Predicted Pr(*overall*=5) - reviewers *v.* non-reviewers (by treatment)

	5-point treatment	**3-point treatment / Control**
Reviewer	0.798***	0.698***
	(0.065)	(0.045)
Not reviewer	1***	0.471***
	(0.000)	(0.329)

The aforementioned findings lead us to conclude that both reviewers and non-reviewers are affected by the 5-point treatment, even though in a different way. More specifically, estimated predictive probabilities signal that non-reviewers are more influenced by the treatment - a further proof that their participation in online rating system would not be enough to correct biases as long as information about prior rating are visible. It seems, though, that the probability of rating 5/5 is higher for reviewers than non-reviewers under the 3-point treatment and in the control group - while it is lower under the 5-point treatment. Nevertheless, non-parametric testing showed that the two groups were statistically comparable across treatments. We do believe that, as in the case of repeat *v.* new customers, a bigger sample would help us in strengthening the results when comparing different sub-segments of the subject pool.

Result 5. *Both reviewers and non-reviewers are influenced by prior ratings. Yet, non-reviewers exhibit a higher influence to excellent prior ratings.*

6 Discussion and Conclusions

The experiment we conducted fits into a stream of recent literature tackling the social and economic issues of UGCs and online rating platforms, particularly aimed at addressing the behavioral attitude of reviewers and assessing the relevance of SIB. If such bias exists, individual ratings are unreliable proxies for the "true" quality of products, especially when distinguishing mediocre from excellent products. Moreover, aggregate ratings (and all related summary statistics) might be the result of herding, thus leading to problems such as bimodal distributions of scores, rating bubbles and the incentive to post fake reviews.

We detached from a few influential studies which conducted in-vivo randomized experiments on online rating systems (Aral and Walker, 2012; Muchnik et al., 2013; Krishnan et al., 2014) and proposed a pseudo-online field experiment in which the novelty of the approach mainly aimed at: i) assessing how, in the case of an experience good, individual rating behavior is affected by information about prior ratings; ii) measuring the reaction of non-reviewers when they are exposed to a rating system, in order to identify any under-reporting bias with respect to

SIB; iii) disentangling and analyzing the rating behavior of repeat customers of the same service with respect to first time visitors. This allowed us to test whether SIB plays a different role depending on whether the service falls more in the realm of experience or search goods. We devoted our attention to tourism services since this sector has been a pioneer in the development of e-commerce and online rating systems and even today it plays a leading role through companies such as Booking, TripAdvisor, Yelp and Expedia.

Our results confirm that SIB is, indeed, a relevant issue in rating systems, in addition to being asymmetric: while subjects herd to the display of information about extremely positive ratings, they are not influenced by information about below-the-average ratings. Moreover, the novelty and relevance of our results is two-fold: first, SIB is not driven by the familiarity of subjects with online rating systems, since reviewers and non-reviewers (the so-called posters and lurkers, respectively) are both affected by prior ratings; hence, SIB is not confounded with any possible under-reporting bias - in fact, when extremely positive ratings are visible, non-reviewers are much more likely to rate accordingly. Second, we show that also repeat customers are significantly affected by prior ratings, as new customers are, albeit with interesting differences: treatment effects on repeat customers are less effective - there is a corrective effect that mitigates the treatment influence. It is likely that this different behavioral pattern has to do with a stronger opinion of customers who chose to stay at the same hotel more than once. Finally, our experimental design was aimed at attracting people who self-select themselves into the market: therefore, any differentiated pattern in rating behavior across treatments cannot be explained through purchasing bias; that is, the buyers' tendency of being more favorable towards the service or the product purchased.

Our results hence provide another voice for supporting reforms in the functioning of online rating systems. In particular, the design of interfaces that provide little information about prior ratings to users when they are asked to review, as well as the experimentation of machine-learning algorithms to estimate and automatically correct biases, might diminish the effect of SIB. Since herd behavior exists, it is important to implement ways to check and correct the bias particularly when a new service is rated. We support the implementation of time windows or numeric thresholds where reviews are collected but not shown to other customers: thanks to this system, a less biased distribution of ratings can emerge, and can then be shown to anyone. These data collection periods can be implemented when a service or product page is just published, but also throughout its online cycle. It is particularly important that online rating systems are designed such that users can escape rating biases. For instance, it is common to show average ratings within soliciting e-mails, when users are reminded to review the service or product they had purchased: this type of information is likely to affect individual rating behavior.

Our work has many limitations, particularly related to the small sample size under analysis. Hence, the collection of more observations, or the repetition of the same experiment on a larger scale would allow us to undertake more convincing and complete within-group analysis. The implementation of new sets of treatments, such as pub-

lic *v.* private posting of the review would allow us to investigate another interesting phenomenon which takes place in online rating platforms: the multiple-audience effect.

Given the relevance of our work for the tourism sector, the expansion of the research scope might work through the replication of the experiment across, and within, different tourism businesses (restaurants, amusement parks, cultural activities), in order to take into account heterogeneous rating distributions, and assess how these affect the overall experience at the destination. In this line of research, since tourism is mainly an experience good, the integration of the dataset with an additional set of explanatory variables, e.g. weather, is of primary urgency, as a way of better explaining and disentangling the drivers of individual rating behavior.

Finally, another possibility is to allow subjects to write a textual review: the rating attitude might change when the consumer is also required to write her opinion about the service or product, on top of the mere rating score. However, another element of complexity (sentiment analysis) would be introduced in the investigation.

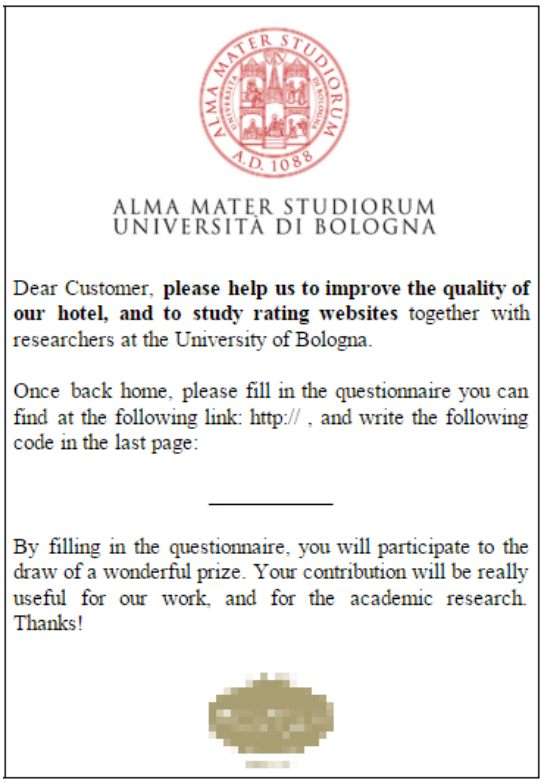
A List of variables

Table 9: List of variables

Variables	Dummy	Definition
Socio-demographic variables		
Female	D	The customer is a female
Age		Customer's age in years
Italian citizenship	D	The customer is Italian
Years of schooling		Customer's years of schooling (from 5, primary education, to 16, university education)
Tourist variables		
Travel type		The customer is travelling in business/ couple/ family/ friends/ solo
Travel length		Number of days spent at the hotel: 2-3 days/ 4-7 days/ more than 7 days
First time destination	D	The customer has never been to the tourist destination before
Repeat customer	D	The customer has already been to the hotel before
Type of stay		Full board/ Full board All Inclusive/ Half Board/ Half Board All Inclusive/ B&B
Period of stay		Period of stay at the hotel: 20/07-09/08 / 10/08-23/08 / 24/08-13/09
Rating behavior variables		
Not reviewer	D	The customer has never written an online review before
Hotel advice		The hotel was advised to the customer by: no one/ family or friends/ advertising/ other
Review read	D	The customer has read an online review of the hotel before booking the stay
Review source		The customer has read an online review of the hotel on TripAdvisor or Yelp/ Facebook or Social networks / Forums/ other
Review influence	D	The customer has been influenced by the review read about the hotel
Other prices	D	The customer looked at other hotel prices before booking her stay
Rating variables		
Overall		Hotel overall rating (from 1, terrible, to 5, excellent)
Excellent rating	D	The hotel overall rating is 5 (excellent)
Sleep quality		Rating for hotel sleep quality (from 1, terrible, to 5, excellent)
Value		Rating for hotel value (from 1, terrible, to 5, excellent)
Service		Rating for hotel service (from 1, terrible, to 5, excellent)
Atmosphere		Rating for hotel atmosphere (from 1, terrible, to 5, excellent)

B Flyer

Figure 3: Flyer handed out to customers



ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Dear Customer, **please help us to improve the quality of our hotel, and to study rating websites** together with researchers at the University of Bologna.

Once back home, please fill in the questionnaire you can find at the following link: http:// , and write the following code in the last page:

_____

By filling in the questionnaire, you will participate to the draw of a wonderful prize. Your contribution will be really useful for our work, and for the academic research. Thanks!

C Questionnaire

Figure 4: First page of the questionnaire - opening layout

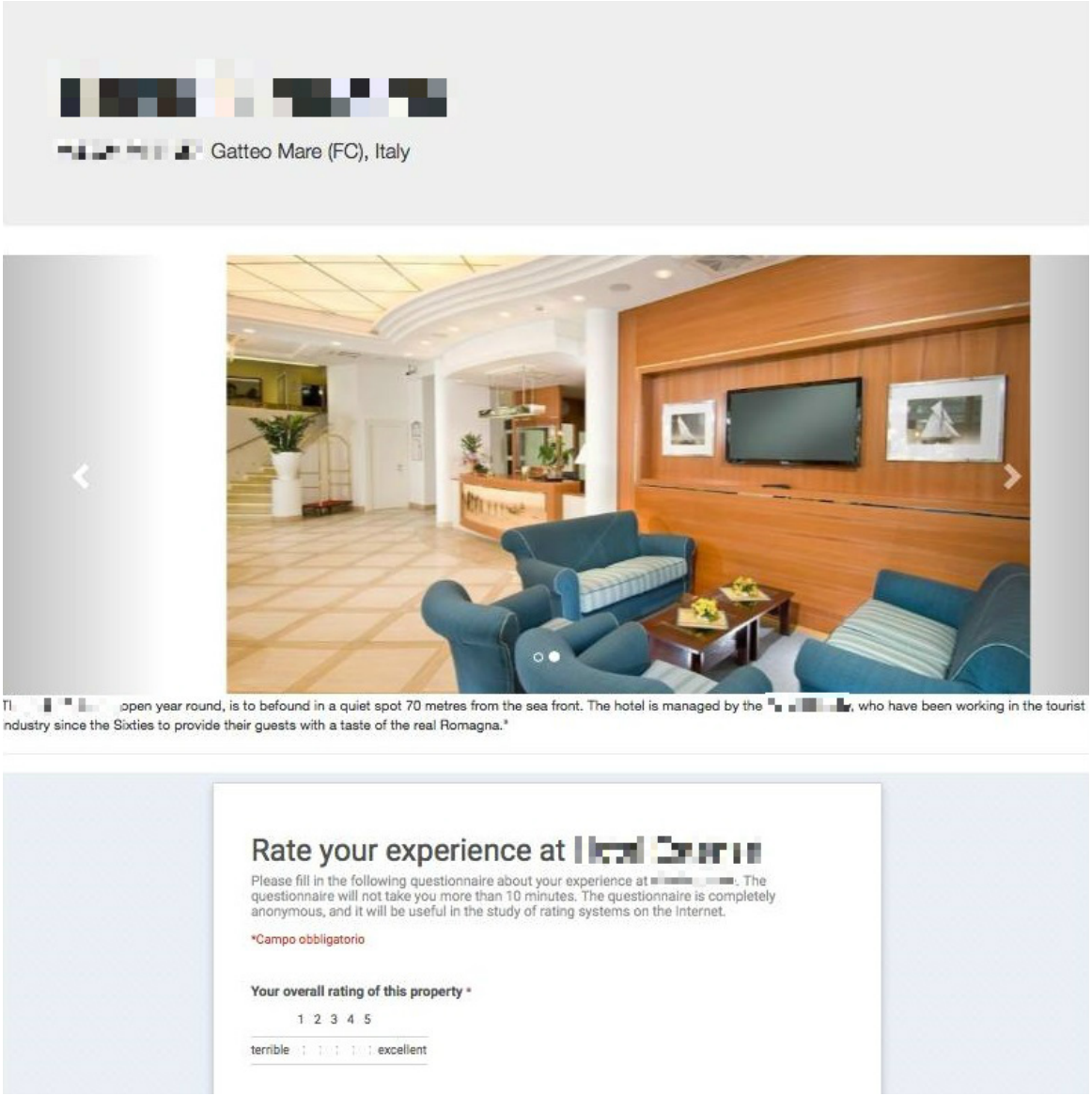
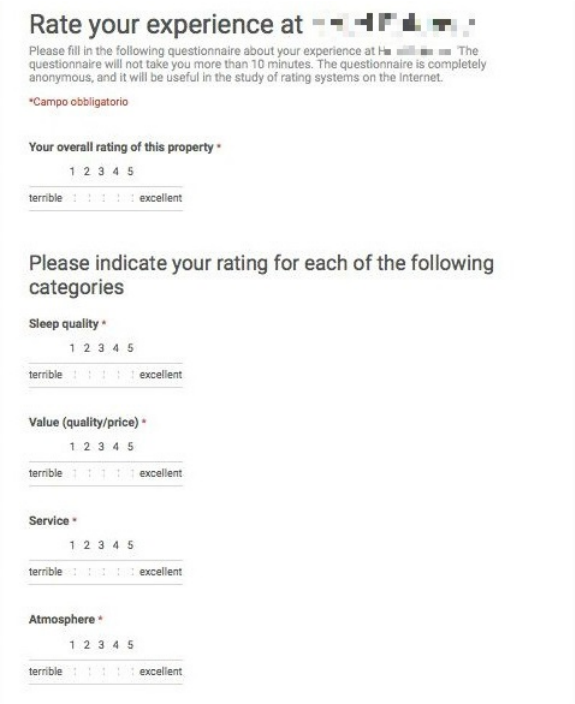


Figure 5: First page of the questionnaire - rating



Rate your experience at [illegible]

Please fill in the following questionnaire about your experience at H[illegible] The questionnaire will not take you more than 10 minutes. The questionnaire is completely anonymous, and it will be useful in the study of rating systems on the Internet.

*Campo obbligatorio

Your overall rating of this property *

1 2 3 4 5

terrible : : : : : excellent

Please indicate your rating for each of the following categories

Sleep quality *

1 2 3 4 5

terrible : : : : : excellent

Value (quality/price) *

1 2 3 4 5

terrible : : : : : excellent

Service *

1 2 3 4 5

terrible : : : : : excellent

Atmosphere *

1 2 3 4 5

terrible : : : : : excellent

Figure 6: Second page of the questionnaire - socio-demographic and tourism questions

Rate your experience at Hotel [redacted]

*Campo obbligatorio

Questionnaire

Gender: *
- Female
- Male

Age: *

Nationality *

Educational qualification: *
- Primary
- Secondary
- High school
- University
- No qualification

What sort of trip was? *
- Business
- Couple
- Family
- Friends
- Solo

How many people were with you during the stay at Hotel [redacted]? *

When did you stay at Hotel [redacted]
Choose the week(s) which include the days of your stay
- 20/7 - 26/7
- 27/7 - 02/8
- 03/8 - 09/8
- 10/8 - 16/8
- 17/8 - 23/8
- 24/8 - 30/8
- 31/8 - 6/9
- 7/9 - 13/9

How much did your trip last overall? *
- 1 day
- 2-3 days
- 4-7 days
- More than 7 days

Was this your first time in Gatteo Mare? *
- Yes
- No

If this was not your first time in Gatteo Mare, how many times have you already been there?

Have you already been at Hotel [redacted] before this trip? *
- Yes
- No

If you have already been at Hotel [redacted] how many times?

Which kind of stay did you have at Hotel [redacted] *
- Full board
- Full board All Inclusive
- Half board
- Half board All Inclusive
- Bed & breakfast

« Indietro Continua »

Figure 7: Third page of the questionnaire - rating attitude questions

Rate your experience at Hotel [redacted]

*Campo obbligatorio

Questionnaire

Do you write restaurant / hotel reviews on web sites and / or social networks: *
- Never
- Rarely
- Once in a while
- Often

Was the hotel recommended to you by anyone? *
- No one
- Family / Friends
- Advertising
- Altro:

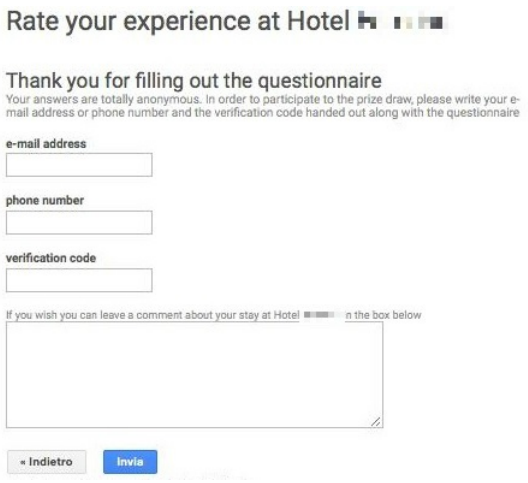
Have you read reviews for Hotel [redacted] on the Internet before booking? *
- Yes
- No

If yes, on which website/s?
- TripAdvisor / Yelp
- Facebook / Social networks
- Forums
- Altro:

If yes, did those reviews influence your choice?
- Yes
- No

Did you have a look at the prices and / or services of other hotels through online web-sites? *
- Yes
- No

Figure 8: Fourth page of the questionnaire - e-mail and verification code



References

Anderson, L. R. and Holt, C. A. (1997). Information cascades in the laboratory. *American Economic Review*, pages 847–862.

Aral, S. (2014). The problem with online ratings. *MIT Sloan Management Review*, 55(2):47.

Aral, S. and Walker, D. (2012). Identifying influential and susceptible members of social networks. *Science*, 337(6092):337–341.

Asch, S. E. (1951). Effects of group pressure upon the modification and distortion of judgments. *Groups, leadership, and men*, pages 222–236.

Asch, S. E. (1955). Opinions and social pressure. *Readings about the social animal*, 193:17–26.

Banerjee, A. V. (1992). A simple model of herd behavior. *The Quarterly Journal of Economics*, pages 797–817.

Banerjee, A. V. (1993). The economics of rumours. *The Review of Economic Studies*, pages 309–327.

Berger, J. and Heath, C. (2008). Who drives divergence? identity signaling, outgroup dissimilarity, and the abandonment of cultural tastes. *Journal of Personality and Social Psychology*, 95(3):593.

Bikhchandani, S., Hirshleifer, D., and Welch, I. (1992). A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of Political Economy*, pages 992–1026.

Bikhchandani, S., Hirshleifer, D., and Welch, I. (1998). Learning from the behavior of others: Conformity, fads, and informational cascades. *The Journal of Economic Perspectives*, pages 151–170.

Bramoullé, Y., Djebbari, H., and Fortin, B. (2009). Identification of peer effects through social networks. *Journal of Econometrics*, 150(1):41–55.

Cai, H., Chen, Y., and Fang, H. (2009). Observational learning: Evidence from a randomized natural field experiment. *American Economic Review*, 99(3):864–82.

Çelen, B. and Kariv, S. (2004). Distinguishing informational cascades from herd behavior in the laboratory. *American Economic Review*, pages 484–498.

Chen, P.-Y., Wu, S.-y., and Yoon, J. (2004). The impact of online recommendations and consumer feedback on sales. *ICIS 2004 Proceedings*, page 58.

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

Chintagunta, P. K., Gopinath, S., and Venkataraman, S. (2010). The effects of online user reviews on movie box office performance: Accounting for sequential rollout and aggregation across local markets. *Marketing Science*, 29(5):944–957.

Cicognani, S. and Mittone, L. (2014). Over-confidence and low-cost heuristics: An experimental investigation of choice behavior. *Economics: The Open-Access, Open-Assessment E-Journal*, 8(2014-36).

Cosley, D., Lam, S. K., Albert, I., Konstan, J. A., and Riedl, J. (2003). Is seeing believing?: how recommender system interfaces affect users' opinions. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 585–592. ACM.

de Albornoz, J. C., Plaza, L., Gervás, P., and Díaz, A. (2011). A joint model of feature mining and sentiment analysis for product review rating. In *Advances in information retrieval*, pages 55–66. Springer.

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

Dellarocas, C., Gao, G., and Narayan, R. (2010). Are consumers more likely to contribute online reviews for hit or niche products? *Journal of Management Information Systems*, 27(2):127–158.

Dellarocas, C., Zhang, X. M., and Awad, N. F. (2007). Exploring the value of online product reviews in forecasting sales: The case of motion pictures. *Journal of Interactive marketing*, 21(4):23–45.

Duan, W., Gu, B., and Whinston, A. B. (2008). Do online reviews matter?an empirical investigation of panel data. *Decision Support Systems*, 45(4):1007–1016.

Duan, W., Gu, B., and Whinston, A. B. (2009). Informational cascades and software adoption on the internet: an empirical investigation. *MIS Quarterly*, pages 23–48.

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):291–313.

Godes, D. and Mayzlin, D. (2004). Using online conversations to study word-of-mouth communication. *Marketing Science*, 23(4):545–560.

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

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

Hu, Y. and Li, X. (2011). Context-dependent product evaluations: an empirical analysis of internet book reviews. *Journal of Interactive Marketing*, 25(3):123–133.

Krishnan, S., Patel, J., Franklin, M. J., and Goldberg, K. (2014). A methodology for learning, analyzing, and mitigating social influence bias in recommender systems. In *Proceedings of the 8th ACM Conference on Recommender systems*, pages 137–144. ACM.

Lai, H.-M. and Chen, T. T. (2014). Knowledge sharing in interest online communities: A comparison of posters and lurkers. *Computers in Human Behavior*, 35:295–306.

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*.

Li, X. and Hitt, L. M. (2008). Self-selection and information role of online product reviews. *Information Systems Research*, 19(4):456–474.

Liao, S. and Chou, E.-y. (2012). Intention to adopt knowledge through virtual communities: posters vs lurkers. *Online Information Review*, 36(3):442–461.

Litvin, S. W., Goldsmith, R. E., and Pan, B. (2008). Electronic word-of-mouth in hospitality and tourism management. *Tourism Management*, 29(3):458–468.

Liu, Y. (2006). Word of mouth for movies: Its dynamics and impact on box office revenue. *Journal of marketing*, 70(3):74–89.

Liu, Z. and Park, S. (2015). What makes a useful online review? implication for travel product websites. *Tourism Management*, 47:140–151.

Luca, M. (2011). Reviews, reputation, and revenue: The case of yelp. com. *Harvard Business School NOM Unit Working Paper*, (12-016).

Luca, M. and Zervas, G. (2013). Fake it till you make it: Reputation, competition, and yelp review fraud. *Harvard Business School NOM Unit Working Paper*, (14-006).

Manski, C. F. (1993). Identification of endogenous social effects: The reflection problem. *The Review of Economic Studies*, pages 531–542.

Moe, W. W. and Schweidel, D. A. (2012). Online product opinions: Incidence, evaluation, and evolution. *Marketing Science*, 31(3):372–386.

Moe, W. W. and Trusov, M. (2011). The value of social dynamics in online product ratings forums. *Journal of Marketing Research*, 48(3):444–456.

Muchnik, L., Aral, S., and Taylor, S. J. (2013). Social influence bias: A randomized experiment. *Science*, 341(6146):647–651.

Ong, L. S., Tan, J. H., et al. (2015). Sense and sensibility of ownership: Type of ownership experience and valuation of goods. *Journal of Behavioral and Experimental Economics*, 58:171–177.

Park, D.-H., Lee, J., and Han, I. (2007). The effect of on-line consumer reviews on consumer purchasing intention: The moderating role of involvement. *International Journal of Electronic Commerce*, 11(4):125–148.

Racherla, P. and Friske, W. (2012). Perceived usefulness of online consumer reviews: An exploratory investigation across three services categories. *Electronic Commerce Research and Applications*, 11(6):548–559.

Schlosser, A. E. (2005). Posting versus lurking: Communicating in a multiple audience context. *Journal of Consumer Research*, 32(2):260–265.

Senecal, S. and Nantel, J. (2004). The influence of online product recommendations on consumers online choices. *Journal of Retailing*, 80(2):159–169.

Smith, D., Menon, S., and Sivakumar, K. (2005). Online peer and editorial recommendations, trust, and choice in virtual markets. *Journal of Interactive Marketing*, 19(3):15–37.

Sparks, B. A. and Browning, V. (2011). The impact of online reviews on hotel booking intentions and perception of trust. *Tourism Management*, 32(6):1310–1323.

Vermeulen, I. E. and Seegers, D. (2009). Tried and tested: The impact of online hotel reviews on consumer consideration. *Tourism Management*, 30(1):123–127.

Viglia, G., Furlan, R., and Ladrón-de Guevara, A. (2014). Please, talk about it! when hotel popularity boosts preferences. *International Journal of Hospitality Management*, 42:155–164.

Wang, T., Wang, D., and Wang, F. (2014). Quantifying herding effects in crowd wisdom. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1087–1096. ACM.

Wu, F. and Huberman, B. A. (2008). How public opinion forms. In *Internet and Network Economics*, pages 334–341. Springer.

Xiang, Z. and Gretzel, U. (2010). Role of social media in online travel information search. *Tourism management*, 31(2):179–188.

Yacouel, N. and Fleischer, A. (2012). The role of cybermediaries in reputation building and price premiums in the online hotel market. *Journal of Travel Research*, 51(2):219–226.

Zhang, K. Z., Lee, M. K., and Zhao, S. J. (2010). Understanding the informational social influence of online review platforms. In *ICIS*, page 71.

Alma Mater Studiorum - Università di Bologna
DEPARTMENT OF ECONOMICS

Strada Maggiore 45
40125 Bologna - Italy
Tel. +39 051 2092604
Fax +39 051 2092664
http://www.dse.unibo.it