On the Supply of Online Reviews: Volume, Valence, and Quality

A DISSERTATION SUBMITTED TO THE

FACULTY OF THE UNIVERSITY OF

MINNESOTA BY

Zhihong Ke

IN PARTIAL FULFILLMENT OF THE

REQUIERMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

Dr. De Liu (Adviser)

May 2019

Acknowledge

First and foremost, I wish to express my most sincere gratitude to my adviser, Dr. De Liu. His continued support and guidance has helped and directed me through my Ph.D. studies and made this dissertation possible. His extraordinary passion and expertise in research inspired and will continue to inspire me to explore truth and create value in my academic career. I am very fortunate to have him as my adviser.

I would also like to sincerely thank Dr. Alok Gupta and Dr. Yuqing Ren for their continued support throughout the past five years. The constructive feedback and insights they have patiently offered benefit this dissertation and my research at large. Their patience, dedication, and respect has inspired me to become the caliber of educator as they are.

I appreciate the help and support everyone in the Department of Information and Decision Sciences has given me. The conversations and laughs we have shared in the past five years are some of the most precious memories of my life.

I am indebted to my sister, Zhiwei Ke, for her unconditional support and encouragement which has shaped who I am, and my brother, Zhijun Ke, who shares my responsibility of taking care of my parents. I also would like to thank my parents, Guifang Zhang and Changming Ke, for their love and support. Last but not least, I am thankful for my husband, Mitch Kostelecky, and the Kostelecky family for their loving support and for making me feel at home in the United States of America.

Table of Contents

List of Tables

List of Figures

Introduction

UGC platforms such as Wikipedia, online review platforms, online Question and Answer forums have permeated many economic and societal activities and established their important roles in our lives. However, the secrets of enabling and sustaining the supply of UGC have only received limited attention. Needless to say, without a steady supply of quality content, UGC would not persist.

My dissertation focuses on the supply of online reviews/ratings – a dominant source for consumers' decision making. Volume, valence, and quality of online reviews are inter-related aspects. Accordingly, this dissertation consists of three chapters focusing on the three aspects. The first chapter explores *whether friend contribution can be used to motivate users to write more and better reviews*. The second chapter examines *whether online ratings are robust to rating aberrations such as random ratings and fake ratings*. The third chapter *develops a measurement for assessing review quality and investigate the relationship of helpfulness votes (a commonly used proxy for review quality) with review quality*.

All three chapters are related to fundamental issues in the supply of online reviews. As is well known, consumers rely on online reviews for a wide range of decisions, ranging from buying goods, dining, professional services, and so on. However, online review platforms face challenges in ensuring the volume and quality of user-generated reviews. Noticing these gaps in the literature, I examine the production and supply of online reviews from three perspectives. In the first chapter, I study the potential of using friends' reviews to nudge users of online reviews platforms to write more and higher quality reviews. In the second chapter, I examine the dynamics of online review valence and volume. In the third chapter, I develop a measurement for assessing the quality of online reviews. Using this measurement, I aim to understand whether/how helpfulness votes of online reviews deviates from content quality.

The first chapter is motivated by whether the friend effect exists in social-networked generation of online reviews. While “friend effects” have been found in a number of other settings such as product adoption, so far there is a dearth of empirical evidence on whether the same holds for social-networked generation of online reviews, where review generation might face a natural identity competition.

To fill this research gap, I attempt to answer the following question: how friend contribution, in the form of reviews written by online friends, can be a way of motivating users to write more and higher-quality reviews? Noting the public-good nature of online reviews, I draw upon theories of pure altruism and competitive altruism to understand the effects of friend contribution on the quantity and quality of review provision. I test our hypotheses using data from Yelp, and find positive effects of friend contribution. Users are three times more likely to provide a review after a friend has written one, and this effect cannot be solely explained by homophily. Furthermore, reviews written after a friend’s review tend to be of higher quality. Interestingly, I also find the effects of friend contribution is greatly shaped by users’ tenure and status. New non-elite users – i.e. those who have a relatively short tenure and have not achieved an elite status – are mostly affected by the contribution of their elite friends, whereas old non-elite users are mostly affected by their non-elite friends. New elites are affected by all types of friends, whereas old elites are only affected by old non-elite friends. These findings hold important implications for online review platform designers and marketers.

The second chapter is motivated by the question: are online rating systems reliable and robust? I examine the ebb and flow (i.e., new ratings deviate from the long-term mean in the opposite direction of prior ratings) of online ratings with a focus on whether aberrations (i.e., ratings deviate from the long-term mean) in online ratings cast a long shadow in subsequent ratings. The study consists of two complementary pieces: first, I use econometric methods to analyze a panel of 8,943 Yelp restaurants in the state of Washington with a focus on the dynamics of their rating valence as a function of recent rating aberrations. Second, to verify the findings from the observational study and to establish a causal relationship between rating aberrations and subsequent changes in rating valence, I design and analyze two randomized online experiments using joke ratings.

Overall, the studies show that aberrations in online ratings do not appear to cast a long shadow on the valence of future ratings and online rating systems are fairly resilient to aberrations. Furthermore, I also show that, unlike a previous finding by (Muchnik et al. 2013) in the context of popularity voting, the correction strength, i.e., the amount of correction (new ratings relative to recent rating aberration), for positive aberrations is on par with those for negative ones. These findings add to the nascent literature on the reliability of online ratings and hold important implications for consumers and online rating platforms.

The third chapter is motivated by the reliance of helpfulness votes as a proxy for review quality. I used Delphi method for ontology design to develop a measurement for review quality and then investigates the relationship between helpfulness votes and review quality. The measurement I develop for review quality includes five dimensions: relevant, trustworthy, comprehensive, well-written, and timely. Applying this measurement to Amazon reviews, I find that consumer-voted helpfulness is a poor indicator of review quality. Review length, rating, and the presence of photos all contribute to the difference between review quality and helpfulness votes. Specifically, shorter and negative reviews tend to receive more helpfulness votes relative to their review quality, and such an inflation tends to occur among low-quality reviews. Interestingly, photo-augmented reviews tend to receive fewer helpfulness votes relative to review quality, and this deflation tends to occur among high-quality reviews. These findings hold important implications for research and design on online reviews and beyond.

The first chapter is a collaboration between my adviser (Dr. De Liu), Dr. Alok Gupta, Dr. Daniel Brass, and myself. The second chapter is a collaboration between my adviser, Dr. Alok Gupta, and myself. The third chapter is a collaboration between my adviser, Dr. Gediminas

Adomavicius, and myself. To acknowledge their contributions, I use "we" to refer to all aforementioned coauthors in each chapter. Next, I will discuss the chapters one by one.

Chapter 1. The Effect of Friend Contribution on Online Review Provision

1. Introduction

User-generated online reviews of products and services have become a dominant source of information for consumers. A 2016 report by BrightLocal concluded that an overwhelming 91 percent of consumers reported reading online reviews, and among them, approximately 93% said that their buying decisions were influenced by online reviews (BrightLocal 2016). Online reviews play an important role in reducing information asymmetry between consumers and vendors (Pavlou et al. 2007), and increasing consumer confidence in decision making (Kumar and Benbasat 2006). Research indicates a consistent positive relationship between the volume of online reviews and sales (e.g., Dellarocas et al. 2007; Duan et al. 2008; Forman and Ghose 2008).

However, very few ordinary consumers write online reviews (Fortune 2016; Goes et al. 2016). Studies estimate that only 1 percent of consumers have ever written an online review (Anderson and Simester 2014; Yelp 2011). As a result, many products and services receive no or very few reviews. To address this issue, online review platforms (ORPs) – platforms that provide user-generated online reviews – have used several approaches to motivate consumers to write online reviews. One category of approaches is to offer tangible rewards, such as financial incentives, coupons, and discounts, for writing a review.[1] Recent research suggests that such tangible rewards are effective, but often have downsides such as resulting lower quality reviews and eroding consumer trust (Burtch et al. 2018; Ghasemkhani et al. 2016; ReviewMeta.com 2016; Stephen et

---

[1] For example, Epinions employs a revenue sharing strategy with reviewers to encourage review generation. Amazon once offered free products to top reviewers and allowed product owners to offer free or discounted products to reviewers in exchange for their reviews, but discontinued this practice under criticism.

al. 2012). Another category is to use social nudges. For example, the online review literature has studied the effect of highlighting aggregate peer contribution (e.g., “2322 users have recently contributed online reviews”) as a form of social nudge (Burtch et al. 2018; Chen et al. 2010). So far, the literature has not explored another form of social nudge – highlighting a specific online friend’s contribution (e.g., “[your friend] has written this review”).

Showing users their friends’ behaviors is a powerful social nudging and has been shown to work in several other online contexts, including adoption of paid music services and products (Bapna and Umyarov 2015; Zhang et al. 2014), music consumption (Dewan et al. 2017), information diffusion (Bakshy et al. 2012), store check-ins (Qiu et al. 2016), and peer-to-peer lending (Liu et al. 2015). One may automatically assume that, by analogy, showing an online user reviews written by their friends must increase the user’s chance of offering a review. We argue that this may not be true for two reasons. First, online reviews are considered public goods. As with other public goods, contribution to online reviews suffers from free-riding: one may be less inclined to contribute when others have contributed. This substitution effect between peer contributions can be stronger when the peer is an online friend. Studies of online communities have shown that online friends often form around similar interests and opinions (Lee et al. 2016; Underwood and Findlay 2004), thus a friend’s review may discourage a user from contributing to the same since the latter is more likely seen as redundant. Second, writing an online review involves significant effort. Friends on ORPs, which in many cases are considered weak ties (Bond et al. 2012; Mesch and Talmud 2006), may not be strong enough to have a measurable impact on such a costly behavior as writing an online review.

A well-known online review platform, Yelp, has cultivated a community of volunteer reviewers on its social platform, and routinely encourages users to make online friends and show them reviews contributed by their online friends. Yelp seems to enjoy a tremendous success in terms of the quantity and quality of reviews produced by its users. However, it is far from clear whether the quantity and quality of online review contribution have anything to do with exposure to friend-contributed reviews. Not only the extant literature on online reviews has not drawn any conclusion on this issue, there does not appear to be an accepted wisdom in the industry either. Yelp appears to be an exception rather than a norm in leveraging the social nudge of friend contribution. Some ORPs, including Foursquare, Zomato, and eBay, also support social networking among their users, but have not leveraged friend contributions. Many other ORPs simply do not support social networking among users.

Intrigued by the aforementioned gap in understanding, we ask the following questions in this research: *Can friend contribution, in the form of reviews written by online social network friends, motivate users of online review platforms to write more and better quality online reviews*? *Furthermore, what type of online friends affect a user's likelihood of contribution?*

To answer the above questions, we draw from theories of public goods to develop and test hypotheses about the likely effect of friend contribution on review quantity and quality. We then test these hypotheses using a unique panel dataset on Yelp users. By observing the reviewing behaviors of a panel of 2,923 users over a set of 8,289 restaurants in the state of Washington over 36 weeks, we examine a user's likelihood of reviewing a restaurant in a particular week increase with the number of friend reviews for the same restaurant in preceding week, while controlling for the potential homophily among friends. To examine the heterogeneous effect of friend contribution, we classify users into four types based on tenure and status: old elites (i.e., those who have a relative long tenure on the platform and have been recognized as an elite reviewer due to their stellar contribution to the platform and the community), new elites, old non-elites, and new non-elites. To examine the effect of friend contribution on review quality, we focus on cases where the user wrote a review and examine relationship between the quality of the review and the number of preceding friend reviews. We measure review quality in two ways: votes received by the review and a separate evaluation conducted on Amazon Mechanical Turk (AMT) (henceforth *Turker evaluation*).

Our research makes several contributions to the literature and practice. First, we contribute to the literature of online reviews by finding a new social nudge using friend contribution in the form of highlighting reviews written by one's friends. Friend contribution can simultaneously increase the quantity and quality of online reviews.

Second, we contribute to the literature on social influence by studying the effect of friend contribution in a new context of online reviews. Our results suggest that although friend reviews may be strong substitutes for each other, the overall effect of friend contribution is positive on both review quantity and quality.

Our result on review quality is novel, as the literature on intervention for reviewing has not considered quality. Finally, we uncover a previously unknown heterogeneity in friend effect: the types of friends a user responds to depend on her and her friends' status and tenure.

2. Related Literature

2.1. Provision of Online Reviews

The online review literature has approached the provision of online reviews from two aspects: valence (what rating to give) and volume/quantity (whether to provide a review). The literature on antecedents of the review valence has examined several factors, including the role of peer ratings (Lee et al. 2015; Ma et al. 2013; Sridhar and Srinivasan 2012) and friend ratings (Lee et al. 2015; Wang et al. 2018). For example, Wang et al. (2015) show that users tend to give similar ratings as those of their friends in book reviews. We note that insights on review valence do not address the question of how to motivate more and higher quality online reviews, which is the focus of our study.

Dichter (1966) proposes a model of online reviews that identifies four main types of motivations for writing online reviews: message involvement as a result of advertisements and public relations, product involvement as a result of particularly positive and negative product experiences, self-involvement as a result of users' need to enhance themselves in front of an audience, and other involvement as a result of a desire to help others make a better decision. Supporting this model, existing research has shown that the volume of online reviews is a function of the characteristics of the products/services (Dellarocas et al. 2010), consumers' consumption experience (Dellarocas and Narayan 2006), and reviewer characteristics (Goes et al. 2014; Moe and Schweidel 2012). Within the stream of research on the review volume, we are most closely related to a few recent studies that examine the effect of a few interventions aimed at increasing the volume of online reviews (Burtch et al. 2018; Chen et al. 2010). Chen et al. (2010) demonstrate that, after being shown the median user's review contribution, users below (above) the median increase (decrease) their contribution to online reviews. Burtch et al. (2018) show that financial incentives can increase review volume, showing users the number of peer contributions can increase review length, and the two can be combined to achieve both benefits. We add to this stream of research by studying a different form of social nudge, friend contribution.

Extent literature on review quality primarily focuses on what kinds of online reviews are considered more "helpful" by consumers. Numerous research studies examine the association between features of a review's textual content and helpfulness votes (Cao et al. 2011; Ghose and Ipeirotis 2011; Mudambi and Schuff 2010; Yin et al. 2014). In this literature stream, a growing number of studies examine contextual factors, such as the product type and reviewer characteristics (Lu et al. 2010; Mudambi and Schuff 2010). The review quality literature has paid little attention to the issue of how to promote review quality.

2.2. Social Influence in Other Online Contexts

In a broad sense, our research is related to a few literature streams that examine the role of social influence in user-generated content. One stream of research investigates the effect of peer influence on voluntary contributions to "electronic communities of practice", such as online discussion forums, Q&A forums, and knowledge-sharing communities (Wasko et al. 2009; Wasko and Faraj 2005). In these online communities, contributions tend to be directed toward specific members of the community (e.g., answering another user's question or participating in a threaded discussion). Hence, researchers have used social-exchange-based theories, such as reciprocity and generalized reciprocity, to explain users' contribution behavior and the role of peer influence (Jabr and Mookerjee 2014; Nam et al. 2009; Xia et al. 2011). In contrast, online review platforms represent a different environment, where user-generated content is considered a contribution to public goods rather than a form of exchange with certain members of the community. Therefore, insights from electronic communities of practice may not directly apply to online reviews.

The social influence literature has also studied the effect of exposing users to their friends' behaviors in various online environments, including adoption of paid music services and products (Bapna and Umyarov 2015; Zhang et al. 2014), music consumption (Dewan et al. 2017), information diffusion (Bakshy et al. 2012), store check-ins (Qiu et al. 2016), and peer-to-peer lending (Liu et al. 2015). Most of these studies is concerned with private behaviors (e.g. consumption) of users, which have a different set of motivations from contributions to a public online review platform. For example, different from previous study contexts, an online review contributor may be motivated to maintain a distinctive online identity by posting novel online reviews. Hence, previous insights on the effect of friend influence may not generalize to our setting.

Finally, a stream of research examines a different kind of social influence: the role of audience size on the voluntary contribution of information goods (e.g. Wikipedia pages and tweets) (Huang et al. 2016; Rui and Whinston 2012; Wang et al. 2017; Zhang and Zhu 2011). These studies find that the bigger the audience, the more contributions.

3. Theoretical Background and Hypotheses

We consider an ORP that facilitates an online community among its users. We use the term "*user*" to refer to a registered member of the ORP. Once a user becomes a registered member, the user can have a profile, contribute reviews, and interact with other members using tools provided by the platform, such as comments, votes, and likes. In addition, the ORP allows users to become "online friends" with each other. Once users become online friends, they can receive updates on each other's activities, such as their reviews and photos. Prior research characterizes such "online friends" as those who share personal interests, and use electronic connection and communication as a primary form of interaction with each another (Dennis et al. 1998; Hiltz and Wellman 1997; Ridings and Gefen 2004).

To encourage contributions and community building, the ORP also recognizes the most outstanding members in terms of the number and quality of contributions and engagement in the community (e.g., participating in community activities and events, and interacting with other members). Such recognitions are often selected by the community, e.g., based on peer nominations and helpfulness votes of reviews. Such community-based recognitions are present in many online communities. For example, Yelp has an elite program, which awards elite badges to members who contribute the most high-quality reviews and do the most to support the community of reviewers each year. In an online Q&A community, Stack Overflow, members making extraordinary contributions (based on peer votes) are awarded badges. Such recognition can lead to tangible benefits for the recognized individuals. For example, business owners often offer them free tasting in exchange for their honest reviews to leverage their influence among the community of reviewers and consumers.

To explain the effect of friend contributions on users' contribution behavior, and noting that online reviews are a form of public goods, we draw upon two high-level theories on why people contribute to public goods, namely the *pure altruism theory* and the *competitive altruism theory*. We next briefly discuss these theories and their implications for the friend effect. We note that the pure altruism theory is quite similar to the other-involvement motivation in Dichter's model (See footnote 3 of Dellarocas et al. 2010). The *competitive altruism theory* explains why contributions to online reviews can be motivated by the pursuit of status and superiority, which is a form of self-involvement in Dichter's model. Online reviews may also be a result of productive involvement, when users have had a particular positive or negative product experience, or message involvement, when users are stimulated by communication such as other reviews (Dellarocas and Narayan 2006). We do not consider the latter two kinds of motivations because they are not relevant for explaining the role of friend contributions.

3.1. Friend effect from the perspective of pure altruism

The theory of *pure altruism* postulates that individuals contribute to public goods because they genuinely care about others' payoff. Pure altruism has received support from several laboratory experiments and field findings (Andreoni and Miller 1996; Harbaugh et al. 2007). Pure altruism is relevant in online reviews because prior research has suggested one of the reasons people write online reviews is because they have a genuine desire to help others make a better purchase decision (Dellarocas et al. 2010; Dellarocas and Narayan 2006). One implication of pure altruism is that people will increase their contribution if it yields higher payoffs for others, such as when the benefiting group or "audience" is larger or when one's contribution is more distinctive. The "audience" effect is supported by lab experiments (Isaac et al. 1994) and a recent natural experiment on Wikipedia (Zhang and Zhu 2011). Prior research has also shown that movie raters contribute more when they are told their contribution is distinctive (Chen et al. 2010; Ludford et al. 2004).

Because friends on online review platforms are formed around shared interests, they tend to hold similar opinions (Dey 1997; Moretti 2011; Sinha and Swearingen 2001). Therefore, seeing a review contributed by an online friend may cause a user to consider her contribution redundant. By the argument of pure altruism, when a user perceives her contribution to be less distinct and have less value to others, her incentive to contribute decreases. Thus, she is less likely to follow a friend review with her own.

3.2. Friend effect from the perspective of competitive altruism

The *competitive altruism theory* of public goods provision, which originated from social psychology, offers an alternative account of motivations behind altruistic contributions. The central tenant of competitive altruism theory is that individuals may behave altruistically to compete for status in the community, which provides them selective benefits in the long run (Hardy and Van Vugt 2006; Willer 2009).

The competitive altruism theory has three key elements. First, altruistic contributions lead to status. That is, the community grants high status to individuals who have made extraordinary altruistic contributions. Second, high-status altruists are rewarded in the long run. For example, high-status altruists may enjoy an advantage in future interactions, because their altruistic actions signal to potential interaction partners their competence and commitment (Hardy and Van Vugt 2006). The community may also choose to reward individuals who display the greatest concerns and commitment to the group (Willer 2009), such as by offering them perks, access to resources (e.g., invitation to exclusive parties and events), and power within the community. Third, when status is at stake, individuals will compete with each other in altruistic contributions (Hardy and Van Vugt 2006; Lampel and Bhalla 2007). The literature has confirmed that status competition can provide strong motivations for voluntary giving (Constant et al. 1996; Donath 2002; Jones et al. 1997) and voluntary contribution in online communities (Levina and Arriaga 2014; Wasko and Faraj 2005). Though the competitive altruism theory is still relatively new, its description is consistent with the existing research on motivations for online reviews: research has suggested that online reviews are often motivated by the pursuit of attention, status, and superiority (Huang et al. 2016; Rui and Whinston 2012; Wang et al. 2017; Zhang and Zhu 2011).

One of the predictions of the competitive altruism theory is that status-striving works better when users are well embedded in a social network (Anderson and Kilduff 2009) because it is easy for a user's contributions to go noticed (Anderson and Shirako 2008). A user's social network friends are more helpful for her status advancement because friends, due to their shared interests and mutual appreciation, are more likely to value and endorse each other's contribution. Therefore, status-seeking altruists will find it more advantageous to impress their friends, such as by following up a friend review with the user's own. Moreover, a friend review also signals the interests on the topic among friends. There is a natural advantage for a user to contribute to a topic that interests her intended audience.

Existing research shows that individuals will compete for status in ways that suggest a high level of competence, generosity, and commitment to the group (Anderson and Kilduff 2009). When a competitive altruist follows up on a friend review, she would strive to produce a high-quality review to impress her friend. Therefore, competitive altruism also suggests that reviews after a friend contribution are more likely to be of higher quality.

We note that the principle of competitive altruism applies also to users who are not the top most contributors. Within the community of volunteer reviewers, there are also more achievable "local" status and associated benefits. For example, on Yelp, there are events and self-organized social activities among non-elite users. To be considered for these status-based benefits, users must also remain competitive among their circle of friends.

To sum up the views from the two theories, we note that the pure altruism theory suggests a negative effect of friend contribution on review quantity, whereas the competitive altruism theory suggests a positive effect of friend contribution on both quality and quantity. We expect the overall effect of friend contribution on review quantity to be positive for two reasons. First, although friends are expected to have similar interests and opinions, a pure altruist may still find differences in her opinions that allows her to make a distinctive contribution. Second, a high-quality contribution is a combination of endowment and effort. The competitive altruism theory suggests that a friend review can bring out the best in a user by driving more effort in demonstrating one's competence and generosity. Thus, we hypothesize:

H1: *A user's likelihood to review a store increases in the number of recent friends' reviews on the same store.*

H2: *The quality of a user's review of a store increases in the number of recent friends' reviews on the same store.*

3.3. Heterogeneous Effects of Friend Contributions

The competitive altruism theory of the friend effect suggests that users could be differently motivated by friend contributions. First, there may be a higher desire for status recognition when a user is relatively new to the community and has not established herself. Older users may either lose hope of gaining such a recognition, or have already done so and thus are less interested in doing it again. We therefore expect that the friend effect is a function of a user's tenure, with relative new users affected more by friend contributions.

Additionally, attainability of status is also different for users at different levels of the status ladder. Users who are high on the status ladder are closer to the goal of achieving a community-wide recognition, thus are more engaged in status seeking and responsive to friend contributions. Moreover, high-status friends are most helpful in status advancement because, compared to low-status friends, high-status friends are more influential in the community and their evaluation of a user can weigh more in the community's recognition. Therefore, high-status users, especially when they are relatively new, are more likely to respond to high-status friend's contribution. Users who are low on the status ladder has little hope of achieving a community-wide recognition, thus are less motivated by friend contribution, and even if they do, they are more affected by low-status friends who is more helpful in helping them gain "local" status.

In sum, we expect the friend effect to differ by the focal user's status and tenure, and by those of her friends. We note that although the literature on online communities has noted that there are marked difference in contribution frequency and social connectivity between elite and non-elite users (Levina and Arriaga 2014), and between new and old users (Danescu-Niculescu-Mizil et al.

2013; Levina and Arriaga 2014; Ren et al. 2007), the literature has not explored the heterogeneous effect of social influence by users status and tenure. Our next hypothesis explores such heterogeneous effect on the quantity of contributions.[2]-Thus, we hypothesize:

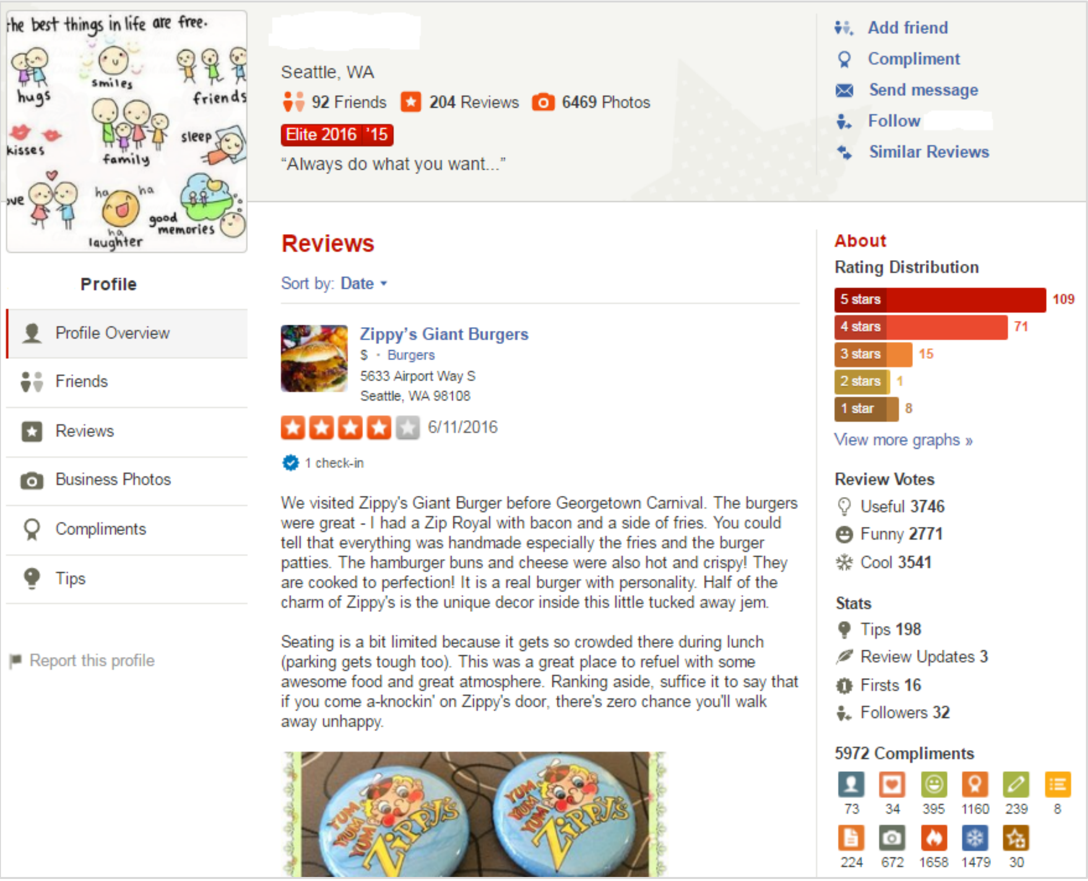
H3: *The effects of friend contribution on a user's likelihood to review a store differ by the user and her friend's tenure and status.*

4. Research Context and Data

We collected our data from Yelp, one of the largest and most successful online review platforms in the world. Yelp operates as a platform for user-generated reviews for local businesses such as restaurants and schools. As of 2017 it had 141 million monthly visitors and 148 million reviews. Yelp has a prominent social networking feature. Each registered Yelp user has a public profile that includes information such as the user's name, location, reviews written, friends, bookmarks, and compliments received (see Figure 1). A user can request to become friends with other users. Once a friend request is confirmed, the friend's reviews will be displayed on the focal user's private homepage under the "friends" section (Figure 2a), and will also be shown on top of the review list on a business's page (Figure 2b). In addition, Yelp users can see stranger reviews on business pages and on their private homepage under the "near you" section (Figure 2a). Yelp does not send push notifications of friend or stranger reviews to users.

---

[2] We do not hypothesize heterogeneous effect because we do not have a strong reason to expect that the review quality after one type of friend to be significantly different that after a different type of friend, once a user decides to write a review after the friend.



Seattle, WA
92 Friends 204 Reviews 6469 Photos
Elite 2016 '15
"Always do what you want..."

Reviews

Zippy's Giant Burgers
$ · Burgers
5633 Airport Way S
Seattle, WA 98108
★★★★☆ 6/11/2016
1 check-in

We visited Zippy's Giant Burger before Georgetown Carnival. The burgers were great - I had a Zip Royal with bacon and a side of fries. You could tell that everything was handmade especially the fries and the burger patties. The hamburger buns and cheese were also hot and crispy! They are cooked to perfection! It is a real burger with personality. Half of the charm of Zippy's is the unique decor inside this little tucked away jem.

Seating is a bit limited because it gets so crowded there during lunch (parking gets tough too). This was a great place to refuel with some awesome food and great atmosphere. Ranking aside, suffice it to say that if you come a-knockin' on Zippy's door, there's zero chance you'll walk away unhappy.

About
Rating Distribution
5 stars 109
4 stars 71
3 stars 15
2 stars 1
1 star 8

Review Votes
Useful 3746
Funny 2771
Cool 3541

Stats
Tips 198
Review Updates 3
Firsts 16
Followers 32

5972 Compliments
73 34 395 1160 239 8
224 672 1658 1479 30

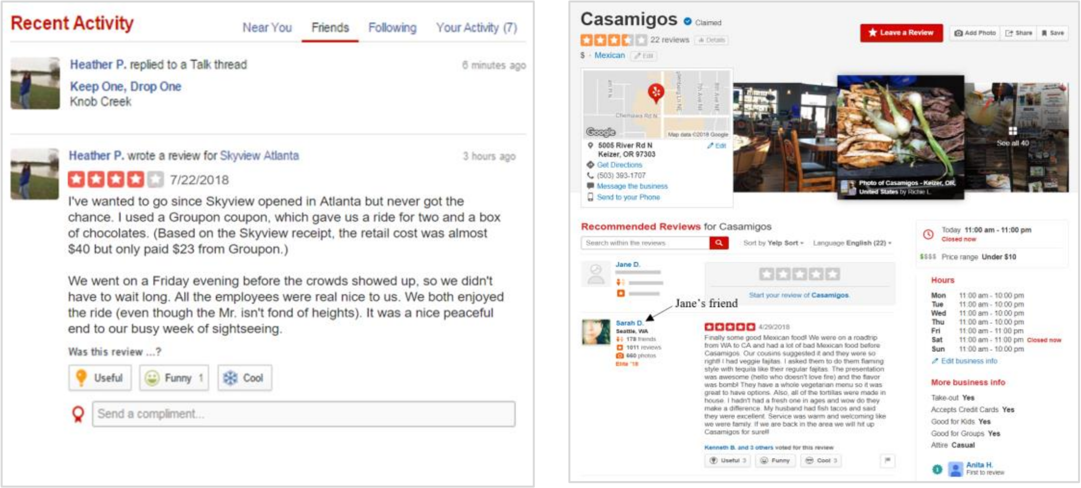
Figure 1.1. An Example of a Yelp User's Public Profile Page

A user can write a review on a business, and/or post a photo of it. They can also vote on existing reviews (no log in required) written by others in terms of its usefulness (useful), humor (funny), or coolness (cool) (Figure 2). Users can also follow other users and send compliments to them. Each year, a Yelp Elite Council selects elite reviewers among nominated candidates based on whether the candidate is a stellar community member and role model.[3] Elite users are honored with a badge on their profile, and receive many perks from Yelp and local business owners.[4]

---

[3] According this blog article (https://www.yelpblog.com/2012/01/what-makes-a-yelper-elite), Yelp does not have a published check list for its Elite criteria. Unofficial sources suggest that elite users are selected based on their last year's review contribution (both quantity and quality), and their engagement with the community, as reflected by their activities such as sending compliments, casting votes, and answering questions. The Elite status is not permanent. A user must earn the Elite badge each year.

[4] Yelp elites receive invitations to exclusive Yelp Elite events (e.g. visits of new business and dinner parties) that are free. They can also bring guests to such events. Because Yelp elite reviewers are very influential, they often get regular invites by businesses for tasting events, Yelp parties, and perks. Business owners are forbidden from soliciting positive reviews.



Heather P. replied to a Talk thread

Keep One, Drop One
Knob Creek

Heather P. wrote a review for Skyview Atlanta

7/22/2018

I've wanted to go since Skyview opened in Atlanta but never got the chance. I used a Groupon coupon, which gave us a ride for two and a box of chocolates. (Based on the Skyview receipt, the retail cost was almost $40 but only paid $23 from Groupon.)

We went on a Friday evening before the crowds showed up, so we didn't have to wait long. All the employees were real nice to us. We both enjoyed the ride (even though the Mr. isn't fond of heights). It was a nice peaceful end to our busy week of sightseeing.

Was this review ...?

5005 River Rd N
Keizer, OR 97303
Get Directions
(503) 393-1707
Message the business
Send to your Phone

Recommended Reviews for Casamigos

Jane D.

Sarah D.
Seattle, WA
178 friends
1011 reviews
660 photos
Elite '18

4/29/2018

Finally some good Mexican food! We were on a roadtrip from WA to CA and had a lot of bad Mexican food before Casamigos. Our cousins suggested it and they were so right! I had veggie fajitas. I asked them to do them flaming style with tequila like their regular fajitas. The presentation was awesome (hello who doesn't love fire) and the flavor was bomb! They have a whole vegetarian menu so it was great to have options. Also, all of the tortillas were made in house. I hadn't had a fresh one in ages and wow do they make a difference. My husband had fish tacos and said they were excellent. Service was warm and welcoming like we were family. If we are back in the area we will hit up Casamigos for sure!!

Kenneth B. and 3 others voted for this review

Today 11:00 am - 11:00 pm
Closed now
$$$$ Price range Under $10

Hours

Mon	11:00 am - 10:00 pm
Tue	11:00 am - 10:00 pm
Wed	11:00 am - 10:00 pm
Thu	11:00 am - 10:00 pm
Fri	11:00 am - 11:00 pm
Sat	11:00 am - 11:00 pm Closed now
Sun	11:00 am - 10:00 pm

Edit business info

More business info

Take-out Yes
Accepts Credit Cards Yes
Good for Kids Yes
Good for Groups Yes
Attire Casual

(a) Friend Review Feeds on a Private Homepage (b). A Friend Review Featured on a Business Page

Figure 2.2. Examples of Friend Reviews on Yelp

We collected data on restaurant reviews in the state of Washington (WA) between March 2013 and November 2013.[5] To obtain a list of users in the WA area who write restaurant reviews, we started with all 551 elite users located in Seattle, WA, then obtained their friend lists, which resulted in 33,815 users. Among the 33,815 elite users' friends, we selected our study sample as those who (a) were located in WA, and (b) had written at least one review on WA restaurants during our study period. The resulting set of 2,923 users accounts for 78% of all users who meet the two criteria, suggesting that we have a fairly comprehensive list of users.[6]

For each user in our study sample, we revisited the user's profile and list of friends every month between March 2013 and April 2014. We also collected all their reviews, bookmarks, and compliments received since March 2012. To ensure that we had complete data on reviews, we separately collected a total of 109,402 reviews on all 8,289 WA restaurants generated during our study period.

---

[5] We picked the WA area because the number of restaurants and the number of reviews per month in this area are close to the average among 21 metropolitan areas featured on the front page of Yelp (Wang 2010).

[6] Our database has collected, over time, all users who have written a review on any of the 8,289 WA restaurants in our dataset, and any users who are either friends, or friends of friends of the 551 elite users. Among all the users at the end of our data collection, a total of 3,748 users who were located in Washington State (WA) and have written at least 1 review on a WA restaurant during our study period.

5. Analysis on Review Quantity

5.1. Dataset, Model, and Variables

To test the effect of friend reviews on review quantity, we constructed a panel dataset at the user × restaurant × period (week) level in the following way. First, we intersected the 8,289 WA restaurants with 2,923 users to obtain 24,228,747 user-restaurant pairs. Among all user-restaurant pairs, 18,387 user-restaurant pairs were *events*, i.e. cases where the user wrote a review for the restaurant during our study period. Because review events are very rare in our data, we followed the suggestion of King and Zeng (2001) to sample all available events (reviews) and a tiny fraction of nonevents (no review), and used weighting to correct the estimated coefficients. Specifically, we kept all events and randomly sampled 5 times of non-events without replacement from all non-events. We then intersected the resulting 110,322 user-restaurant pairs with 36 periods to obtain 3,971,592 user-restaurant-period triples. Finally, we dropped cases where users had already written a review for the given restaurant, and obtained 3,663,479 cases for our analysis.

Our dependent variable, $Review_{ijt}$, is a binary indicator of whether user *i* wrote a review on restaurant *j* in period *t* (i.e., whether user *i* survives in period *t*). Because a user reviews a restaurant at most once, and the panel consists of discrete periods, we adopted a discrete-time survival model for our data, where an event (death) is a review. The discrete-time survival model is equivalent to the logit model as the discrete-time hazard is the conditional odds of dying (i.e. writing a review) at each time given survival up to that point.

A problem with rare-event data is that logit models are known to sharply underestimate event probabilities in samples with less than 200 events (King and Zeng 2001). To correct such a bias, Rare Event logit (ReLogit for short) was proposed by King and Zeng (King and Zeng 2001, 2002). ReLogit is an estimation technique that estimates the same logit model, but with an estimator that gives lower mean square errors for coefficients, probabilities, and other quantities of interest in the case of rare-events data. The number of events in our data (18,387) is not small. Still, as a precaution, we used ReLogit as our main estimation technique, and logit as a backup.

A potential confound of friend effect is homophily – a pair of friends independently chose to review the same restaurant because of their similar preferences. To control for homophily, we followed Wang et al. (2015)'s approach by including reviews written by future friends as a control variable. A future friend is defined as a user who becomes a friend of the focal user in a future period. Because they eventually become friends, they share similar preferences, thus a homophily effect exists. However, because they are not friends at the time of the review, the future friend's review would not have had a friend effect on the focal user. Thus, any effect of a future-friend review is a result of homophily only. If the effect of a current-friend review exceeds that of a future-friend review, we can infer the existence of friend effect beyond homophily.

Formally, we assume the utility for user *i* to write a review on restaurant *j* in period *t*, U_{ijt}, is a function of the number of reviews in period *t-1* written by current friends on restaurant *j* ($CurFrndReviews_{i,j,t-1}$), the number of reviews in period *t-1* written by future friends on restaurant j ($FutFrndReviews_{i,j,t-1}$), the number of new reviews in period *t-1* on restaurant *j* ($NewReviews_{j,t-1}$), additional control variables, and an i.i.d. random component ε_{ijt} with a type I extreme value distribution.

$$U_{ijt} = \beta_1 CurFrndReviews_{i,j,t-1} + \beta_2 FutFrndReviews_{i,j,t-1} + \beta_3 NewReviews_{j,t-1} + \gamma Controls_{i,j,t-1} + \varepsilon_{ijt} \qquad (1)$$

Where $NewReviews_{j,t-1}$ captures the effect of stranger reviews in period *t-1*.

Control variables. We included an extensive list of control variables (see Table 1 for description). We first controlled for several user characteristics. Following Wang (2010), we controlled for the number of compliments sent and received *(#Compliments*), and the number of friends (*Log#Friends*). We used the number of reviews by the user in the last period *(#SelfReviews*) and the number of cumulative reviews by the user up to the last period (*Log#CumSelfReview*) to control for a user's tendency to write reviews. To control for the life cycle of users on the platform, we included tenure on the platform (*LogTenure*). We also controlled for a number of other user characteristics including elite status (*Elite*), gender (*Female*), and estimated income *(CityIncome)*. We inferred gender from the users' reported first names using Behind the Name's database (https://www.behindthename.com) that lists 21,100+ names and their genders.[7] Estimated income was approximated by median household income of the city where the user lives. We used the distance between users and restaurants to capture geographical proximity (*Dist*).

We controlled for a number of restaurant characteristics that may affect a user's review decision, including the restaurant's average rating (*AvgRatingRestaurant*), variance of existing ratings (*AvgVariRestaurant*), cumulative reviews (*Log#CumReviews*), price range (*Price*), whether the restaurant page has been claimed by its owner (*Claimed*), restaurant categories (*11 latent category dummies*), and whether the restaurant was promoted by Yelp at period *t* (*Promoted*). Specifically, we included the restaurant's review valence (*AvgRatingRestaurant*), variance (*AvgVariRestaurant*), and volume (*Log#CumReviews*) because prior research suggested that these affect the quantity of new reviews (Moe and Schweidel 2012). *Price* was coded from levels 1 through 4 based on Yelp reported price ranges ($ to $$$$). We included *Price* because it is likely that dining at expensive restaurants brings more excitement to the user, which in turn increases the likelihood that she will write a review. We controlled for *Claimed* because a claimed store more likely listens to online reviews, which may encourage users to submit reviews. The variable *Promoted* indicates whether the restaurant was featured in the Yelp weekly email to users in period *t*. This variable allows us to control for marketing campaign effects. We obtained 12 latent restaurant categories using Latent Dirichlet Allocation (LDA) – a widely used topic modelling technique. There are over 1,000 distinct restaurant categories in our raw data with each restaurant mapped to one or many of these. It is impractical to include them all in our analysis. Furthermore,

---

[7] For first names that were not in the database or gender ambiguous, we used users' profile photos to identify their gender.

we note that some of the original categories are closely related to each other. For example, “breakfast brunch,” “bakeries,” and “bagels” could belong to the same latent category. LDA allowed us to extract latent restaurant categories (or “topics”) based on word co-occurrences. We first extracted the latent categories, then obtained latent category dummies for each restaurant. Finally, to control for temporal shocks to review quantity, we included month dummies. Table 1 provides summary statistics of the data set.

Table 1.1. Descriptive Statistics of Variables (N = 3,663,479)

Variables	Definition	Mean	Std. Dev	Min	Max
$Review_{ijt}$	Whether user *i* writes a review on restaurant *j* in period *t*: yes 1; otherwise 0	0.01	0.07	0.00	1.00
$CurFrndReviews_{i,j,t-1}$	# current-friend reviews of user *i* on restaurant *j* in period *t-1*	0.00	0.03	0.00	5.00
$FutFrndReviews_{i,j,t-1}$	# future-friend reviews of user *i* on restaurant *j* in period *t-1*	0.00	0.02	0.00	5.00
$NewReviews_{j,t-1}$	# new reviews on restaurant *j* in period *t-1*	0.42	1.06	0.00	38.00
$\#Compliments_{i,t-1}$	# of compliments sent and received by user *i* in period *t-1*	0.11	0.43	0.00	5.38
$\#SelfReviews_{i,t-1}$	# of reviews written by user *i* in period *t-1*	0.21	0.90	0.00	42.00
$Log\#CumSelfReview_{i,t-1}$	Log # cumulative reviews by user *i* up to period *t-1*	3.99	1.35	0.00	7.37
$LogTenure_{i,t-1}$	Log days elapsed since user *i* registered on Yelp up to period *t-1*	7.09	0.52	3.85	8.03
$Log\#Friends_{i,t-1}$	Log (1+ # friends of user *i* in period *t-1*)	3.52	1.08	1.10	7.00
$Elite_{i}$	Whether user *i* is an elite user	0.36	0.48	0.00	1.00
$Female_{i}$	Whether user *i* is female	0.45	0.50	0.00	1.00
$CityIncome_{i}$	Median household income (thousands of dollars) of the city user *i* lives	69.37	13.72	24.49	192.25
$Dist_{i,j}$	Miles between restaurant *j* and the city where user *i* lives	50.35	66.57	0.00	439.94
$AvgRatingRestaurant_{j,t-1}$	Cumulative average rating of restaurant *j* up to period *t-1*	3.59	0.69	0.50	5.00
$AvgVariRestaurant_{j,t-1}$	Variance of cumulative ratings of restaurant *j* up to period *t-1*	1.07	0.30	0.00	2.00
$Log\#CumReviews_{j,t-1}$	Log # cumulative reviews of restaurant *j* up to period *t-1*	3.42	1.24	0.00	7.85
$Promoted_{j,t-1}$	Whether restaurant *j* is promoted in period *t-1*	0.00	0.02	0.00	1.00
$Claimed_{j,t-1}$	Whether restaurant *j*'s business page on Yelp is claimed in period *t-1*	0.66	0.47	0.00	1.00

$Price_j$	Price range of restaurant *j*: 1 - least expensive; 4 - most expensive	1.62	0.56	1.00	4.00

We omit the summary statistics of 8 month dummies and 11 restaurant-category dummies for brevity.

5.2. Main Results on Review Quantity

Prior to estimating the models, we conducted collinearity tests and found no signs of collinearity (VIF < 3). We estimated three models, starting with only control variables, then adding current friends' reviews and new reviews in the last period, and finally adding future friends' reviews. We ran both ReLogit and logit models with weighting adjustments. The results are shown in Table 2. Because the results are consistent between models, we report ReLogit Model 3.

CurFrndReviews has a positive effect ($OR = 2.947$, $p < 0.001$, OR for odds ratio). *FutFrndReviews* also has a positive effect ($OR = 2.18$, $p < 0.001$), but smaller than *CurFrndReviews*'. An F-test comparing the odds ratios for *CurFrndReviews* and *FutFrndReviews* is significantly different ($F = 4.48$, $p = 0.034$), indicating the existence of friend effect beyond homophily. Thus ***Hypothesis 1*** is supported.

Compared with current friend's reviews, stranger reviews (*NewReviews*) have a much smaller effect ($OR = 1.090$, $p < 0.001$). To understand the effect size of *CurFrndReviews*, we computed the predicted probabilities of focal users writing a review when *CurFrndReviews* equals 0 and when *CurFrndReviews* equals 1, holding all other predictors at their means. We find that the probability of writing a review triples when there is a friend review, compared with a stranger review (0.0000223/0.0000076).[8]

Table 1.2. Friend-Review Effect on Review Quantity – Discrete-time Hazard Models

Independent Variables	**ReLogit**			**logit**		
	Model 1	**Model 2**	**Model 3**	**Model 4**	**Model 5**	**Model 6**
	OR	OR	OR	OR	OR	OR
	(SE)	(SE)	(SE)	(SE)	(SE)	(SE)
$CurFrndReviews_{i,j,t-1}$		2.989***	2.947***		2.981***	2.940***
		(0.240)	(0.236)		(0.239)	(0.236)

---

[8] Because we controlled for the total number of reviews in period *t-1*, *CurFrndReviews* = 0 corresponds to the case where all last-period reviews are stranger reviews. We thus interpret the comparison between *CurFrndReviews* = 1 and *CurFrndReviews* = 0 as a friend review versus a stranger review.

NewReviews $_{j,t-1}$		1.091***	1.090***		1.091***	1.090***
		(0.004)	(0.004)		(0.004)	(0.004)
FutFrndReviews $_{i,j,t-1}$			2.180***			2.160***
			(0.253)			(0.251)
#Compliments $_{i,t-1}$	1.361***	1.351***	1.345***	1.361***	1.351***	1.345***
	(0.016)	(0.016)	(0.016)	(0.016)	(0.016)	(0.016)
#SelfReviews $_{i,t-1}$	1.120***	1.121***	1.120***	1.120***	1.121***	1.120***
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Log#CumSelfReview $_{i,t-1}$	1.457***	1.453***	1.454***	1.457***	1.453***	1.454***
	(0.015)	(0.015)	(0.015)	(0.015)	(0.015)	(0.015)
LogTenure $_{i,t-1}$	0.721***	0.718***	0.717***	0.721***	0.718***	0.717***
	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)
Log#Friends $_{i,t-1}$	0.912***	0.913***	0.912***	0.912***	0.913***	0.912***
	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)
Elite $_{i}$	1.998***	1.985***	1.985***	1.998***	1.985***	1.985***
	(0.039)	(0.039)	(0.039)	(0.039)	(0.039)	(0.039)
Female $_{i}$	0.984	0.977	0.977	0.984	0.977	0.977
	(0.015)	(0.015)	(0.015)	(0.015)	(0.015)	(0.015)
CityIncome $_{i}$	1.001+	1.001*	1.001*	1.001+	1.001*	1.001*
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Dist $_{i,j}$	0.983***	0.983***	0.983***	0.983***	0.983***	0.983***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
AvgRatingRestaurant $_{j,t-1}$	1.232***	1.213***	1.212***	1.232***	1.213***	1.212***
	(0.022)	(0.022)	(0.021)	(0.022)	(0.022)	(0.022)
AvgVariRestaurant $_{j,t-1}$	0.643***	0.708***	0.709***	0.643***	0.709***	0.709***
	(0.026)	(0.028)	(0.028)	(0.026)	(0.028)	(0.028)
Log#CumReviews $_{j,t-1}$	1.739***	1.598***	1.597***	1.739***	1.598***	1.597***
	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)
Promoted $_{j,t-1}$	3.038***	2.893***	2.887***	3.008***	2.864***	2.859***
	(0.452)	(0.431)	(0.430)	(0.448)	(0.427)	(0.426)
Claimed $_{j,t-1}$	1.096***	1.112***	1.113***	1.096***	1.112***	1.113***
	(0.021)	(0.021)	(0.021)	(0.021)	(0.021)	(0.021)
Price $_{j}$	1.234***	1.244***	1.243***	1.234***	1.244***	1.243***
	(0.018)	(0.018)	(0.018)	(0.018)	(0.018)	(0.018)
Constant	0.000***	0.000***	0.000***	0.000***	0.000***	0.000***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Month & restaurant category dummies	included	included	included	included	included	included
Log-Likelihood	-	-	-	-199,930	-199,439	-199,416
Pseudo R-squared	-	-	-	0.075	0.078	0.078
N	3,663,479	3,663,479	3,663,479	3,663,479	3,663,479	3,663,479

DV = whether user *i* reviews restaurant *j* in period *t* (*Review*$_{ijt}$). The values in parentheses are standard errors. + p<0.10, * p<0.05, ** p<0.01, *** p<0.001

The effects of most control variables are in the expected directions. *#Compliments* has a positive effect, suggesting that socially active users are more likely to provide reviews. Both *#SelfReviews* and *Log#CumSelfReview* have a positive impact, demonstrating that productive users tend to write more. As one would expect, *Elite* and *CityIncome* have a positive effect, whereas *LogTenure* and *Dist* have a negative effect. Interestingly, *Log#Friends* has a negative impact, suggesting that having many friends who do not review the focal restaurant is negatively associated with the user's probability of reviewing the restaurant. This could be a result of normative influence (Kaplan and Miller 1987; McGrath 1984), i.e. a friends' choice of not reviewing may persuade the focal user not to provide a review either. Consistent with literature (Moe and Schweidel 2012), *AvgRatingRestaurant* and *Log#CumReviews* both have a positive impact, demonstrating that users tend to review highly rated and often-reviewed restaurants. Interestingly, *AvgVariRestaurant* has a negative effect, suggesting that users are less likely to review the restaurants if prior users have very different opinions. Disagreements among prior consumers may prohibit later users from visiting the restaurants, thus, less new reviews are generated. Not surprisingly, *Promoted*, *Claimed*, and *Price* all have a positive impact, suggesting that users were more likely to review promoted, claimed, and pricey restaurants.

5.3. Heterogeneous Effects by User Types

Having established the main effects, we further examined whether/how friend effect differs across types of users. Following prior literature (Crowston and Wei 2006; Dahlander and Frederiksen 2012; Paolillo 2008; Danescu-Niculescu-Mizil et al. 2013; Levina and Arriaga 2014; Ren et al. 2007), we classified users by their status (as indicated by their Yelp elite status) and tenure in the community. Instead of using arbitrary thresholds for the classification, we conducted a k-Means cluster analysis on elite status and tenure, and obtained four clusters, which we labeled as new elites (*NElite*), old elites (*OElite*), new non-elites (*NNonelite*) and old non-elites (*ONonelite*)

(Table 3). In general, a typical new user has been on the platform for fewer than two years, whereas an old user has been on the platform for about 4.5 years.

Table 1.3. Cluster Analysis of User Types

Cluster (Type of Users)	**Cluster Centers**	
	Tenure (days)	**Elite**
New elites	773	1
New non-elites	572	0
Old elites	1,657	1
Old non-elites	1,503	0

We classified current friend reviews using the newly obtained user types for both the focal user and her friends. This resulted in 14 buckets for current friend reviews,[9] which we labeled using the focal users' types and the friends' user types. For example, *NElite* x *OEliteFrnd* represents the number of old-elite friend reviews received by a new-elite focal user. ReLogit results are reported in Table 4, Model 1. We find that all buckets of current-friend reviews have a positive effect except for *NElite* x *ONoneliteFrnd*.

Table 1.4. Further Insights on Review Quantity – Heterogeneous Users

Independent Variables	**Model 1**	**F - tests: Compared with Future Friend Reviews**	
	Odds Ratio (SE)	*F*	*p*
NElite x NEliteFrnd $_{i,j,t-1}$	3.839*** (0.993)	3.79	**0.052**
NElite x NNonEliteFrnd $_{i,j,t-1}$	8.066** (5.603)	3.38	**0.066**
NElite x OEliteFrnd $_{i,j,t-1}$	3.802*** (0.775)	5.42	**0.020**
NElite x ONonEliteFrnd $_{i,j,t-1}$	2.522 (1.493)	0.05	0.830
NNonElite x NEliteFrnd $_{i,j,t-1}$	10.917*** (6.167)	7.66	**0.006**
NNonElite x OEliteFrnd $_{i,j,t-1}$	5.196*** (2.326)	3.40	**0.065**
NNonElite x ONonEliteFrnd $_{i,j,t-1}$	5.872+	0.94	0.332

---

[9] Two of the 16 buckets were dropped because they were empty: new non-elites did not have any new non-elite friends, and old non-elites did not have any new elite friends in our data.

	(5.862)		
OElite x NEliteFrnd $_{i,j,t-1}$	2.654**	0.30	0.587
	(0.802)		
OElite x NNonEliteFrnd $_{i,j,t-1}$	4.684**	1.91	0.167
	(2.483)		
OElite x OEliteFrnd $_{i,j,t-1}$	1.804**	0.89	0.346
	(0.328)		
OElite x ONonEliteFrnd $_{i,j,t-1}$	4.942***	7.84	**0.005**
	(1.294)		
ONonElite x NNonEliteFrnd $_{i,j,t-1}$	78.603***	11.81	**< 0.001**
	(81.098)		
ONonElite x OEliteFrnd $_{i,j,t-1}$	2.920**	0.47	0.494
	(1.130)		
ONonElite x ONonEliteFrnd $_{i,j,t-1}$	7.142***	7.01	**< 0.001**
	(3.059)		
FutFrndReviews $_{i,j,t-1}$	2.216***		
	(0.253)		
NewReviews $_{j,t-1}$	1.090***		
	(0.004)		
Constant	0.000***		
	(0.000)		
N	3,663,479		

DV = whether user *i* reviews restaurant *j* in period *t* ($Review_{ijt}$). The values in parentheses are standard errors. Control variables were the same as the main analysis but omitted for brevity.
+ p<0.10, * p<0.05, ** p<0.01, *** p<0.001.

As in the main analysis, we conducted F-tests on homophily effects (Table 4, col 2). We summarize the findings in Figure 3. The F-tests comparing current and future friend reviews show that new elites are positively affected by reviews from new-elite, new-non-elite, and old-elite friends (F = 3.79, p = 0.052; F = 3.38, p = 0.066; F = 5.42, p = 0.020; respectively). New non-elites are positively affected by reviews from new and old elite friends (F = 7.66, p = 0.006; F = 3.40, p = 0.065; respectively). Old elites are positively affected by reviews from old non-elite friends (F = 7.84, p < 0.005). Old non-elites are positively affected by reviews from new non-elite and old non-elite friends (F = 11.81, p < 0.001; F = 7.01, p < 0.001; respectively). These findings suggest that friend effects differ by the focal user and her friend's status and tenure, thus, ***Hypothesis 3*** is supported.

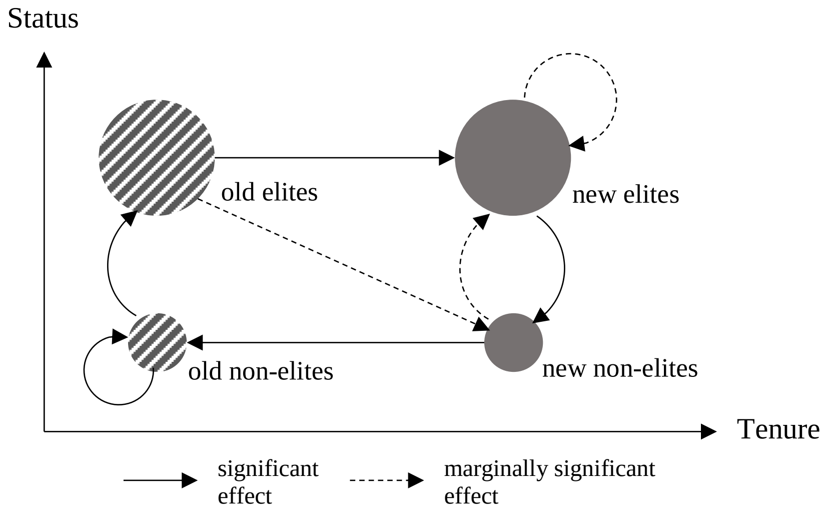


Figure 3.3. Direction of Effects between User Types

Synthesizing our findings, we notice a "tenure effect": there are marked differences in how old and new users choose their sources of friend effect. Specifically, the new elites were open to being affected by nearly all types of friends, whereas old elites were only affected by old non-elite friends, which may reflect the evolution of their commitment and passion towards the communities of reviewers. New non-elites only responded to elite friends whereas old non-elites only responded to non-elite friends. This may reflect whether they were still inspired to climb the status ladder.

In the related vein, there appears to be a "cohort effect": new elites and new non-elites respond to each other' reviews, and old elites respond only to old non-elites. It is interesting that such cohort-based bonds can break through the barrier of social status. However, as time goes on, the cohort effect diminishes for non-elites, as they no longer respond to old-elite friends. In this case, the friend effect falls within the same status.

5.4. Robustness Checks and Further Insights

5.4.1. Are Future Friend Reviews a Good Proxy for Homophily?

In our main analysis, we used future friend reviews as a proxy for homophily between focal users and their current friends, which relied on an assumption that future friends are as similar to the focal user as current friends. In the present section, we validated this assumption in three ways: (a) by directly comparing the focal user's similarities with current and future friends, (b) by restricting both current and future friends to be recent friends to minimize shifts in homophily effects caused by the focal user's evolving preferences, and (c) by comparing effects of friend reviews before and after the friendship formation, which largely holds the homophily effect constant.

Comparing Similarity with Current and Future Friends. A direct way of validating our assumption is to compare the similarity of future friends with that of current friends. We used categories of restaurants in a user's last 5 reviews to capture the user's preferences. For each focal user, we computed, in each period, the similarity between him and 1 randomly selected current friend, and between him and 1 randomly selected future friend based on categories of the restaurants they reviewed. We used two common similarity measures, Jaccard and Cosine similarity.

Table 1.5. Similarities with Current and Future Friends

Variables	Jaccard Similarity				Cosine Similarity			
	Mean	**Std Err**	**t**	**p**	**Mean**	**Std Err**	**t**	**p**
Similarity between focal users and **current friends**	0.785	0.0042	1.025	0.306	0.439	0.0043	-4.539	< 0.001
Similarity between focal users and **future friends**	0.779	0.0040			0.466	0.0042		

We conducted paired t-tests to compare similarities between future friends and current friends. The results of t-tests are reported in Table 5. The Jaccard similarity between focal users and their current friends is not significantly different from that between focal users and their future friends. However, future friends are more similar to focal users than current friends by the Cosine measure, possibly caused by gradual shifts in the focal users' preferences. The effect of this difference, however, was that we had over-estimated homophily, and underestimated friend effect in our main analysis, which implies that friend effect could be even stronger than we reported. Therefore, overall, this analysis suggests that our main finding is robust.

Restricting to New Friends. One potential reason that future friends may not be as similar to focal users as current friends is that these friendships are formed over time and the user's preference shifts over time. To mitigate this concern, we created a subsample in which there was no more than one current or future friend review, and the current and future friends were formed within 90 days of the beginning of the current period. This increases the chance that current and future friends are similar, making future friend reviews a better proxy for homophily.

We used the same model as our main analysis, and the results of ReLogit for this subsample are shown in Table 6, Model 1. Consistent with the main analysis, *CurFrndReviews* and *FutFrndReviews* both have a positive and significant impact ($p < 0.001$). Similarly, the F-test shows that the odds ratios for *CurFrndReviews* and *FutFrndReviews* are significantly different ($F = 4.56$, $p = 0.033$), suggesting the existence of the friend effect.

Table 1.6. Two Robustness Checks

Independent Variables	**Model 1: Homophily Check: Restrict to New Friends**	**Model 2 Lagged Effect of Friend Reviews**
	Odds Ratio	Odds Ratio
	(SE)	(SE)
CurFrndReviews $_{i,j,t-1}$	3.260***	2.495***
	(0.452)	(0.231)
CurFrndReviews $_{i,j,t-2}$	-	2.215***
	-	(0.206)
CurFrndReviews $_{i,j,t-3}$	-	1.926***
	-	(0.209)
NewReviews $_{j,t-1}$	1.092***	1.085***
	(0.004)	(0.005)
FutFrndReviews $_{i,j,t-1}$	2.382***	1.811***
	(0.275)	(0.254)
FutFrndReviews $_{i,j,t-2}$	-	1.839***
	-	(0.277)
FutFrndReviews $_{i,j,t-3}$	-	1.756***
	-	(0.270)
N	3,660,933	3,663,477

We included the same control variables as our main analysis, but omitted them from the report for brevity. Results on control variables are qualitatively the same as our main model. DV = whether user *i* reviews restaurant *j* in period *t* ($Review_{ijt}$). The values in parentheses are standard errors. + p<0.10, * p<0.05, ** p<0.01, *** p<0.001

Future and Current Friends Being the Same Person. To lend further support to our approach, we compared the effect of current and future friend reviews when the future and current friends are actually the same person – i.e., before and after the person forms a friendship with the focal user. Under this stricter condition, the effect of homophily is largely held constant, allowing us to get a more accurate estimation of the homophily effect. We used a matched subsample that consisted of two cases for each user pair (the focal user and an alter), such that the alter wrote a review in period *t-1* in both cases (on different restaurants), but in one case, the alter was the focal user's *future friend* within 90 days, whereas in the other case, the two were *current friends* for less than 90 days. We obtained a total of 84 pairs, which we used to conduct a pairwise comparison.

Because the size of this matched sample is small, we conducted a simple one-way ANOVA analysis. The "current friend" group had a mean *Reviews* of 0.29 (SD = 0.46); The "future friend" group had a mean *Reviews* of 0.09 (SD = 0.29). There is a statistically significant difference between the two groups (*F* (1,82) = 4.97, *p* = 0.028). The pairwise comparison of the two groups reveals that *Reviews* is higher when there is a review by a current friend than when there is a review by a future friend with the same identity, suggesting the existence of friend effect.

5.4.2. Do Friend Reviews Have Lagged Effects?

In our main analysis, we only considered the effect of friend reviews in the last period. If friend reviews from previous periods can also impact the focal user's reviewing decision, our estimation may be biased. To check the robustness of our results against the lagged effect of friend reviews, we added lagged friend reviews in periods *t-2* and *t-3* ($CurFrndReviews_{t-2}$, $CurFrndReviews_{t-3}$, $FutFrndReviews_{t-2}$, $FutFrndReviews_{t-3}$) and reran the analyses.

ReLogit results are reported in Table 6, Model 2. We find that *CurFrndReviews* and *FutFrndReviews* in the last 3 periods remain significant. The F-tests comparing current and future friend reviews show that the odds ratios for current and future friend reviews in period *t-1* are statistically different (F = 3.36, p = 0.067), but those in periods *t-2* and *t-3* are not (F = 1.07, p = 0.301; F = 0.23, p = 0.628; respectively). This result suggests that focal users are only influenced by friend reviews in the last period, not before, and that our main finding is robust after accounting for lagged friend reviews. This also confirms our initial heuristic of using a one-week window for capturing users' response to friend reviews, indicating that it does not take very long for the focal users to react to friend reviews. This is likely because users either take notice of friend reviews and act on it quickly, or they may never act on the friend reviews.

5.4.3. Can the Observed Effect Be Explained by Friends Going to Restaurants Together?

One concern could be that friends tend to go to restaurants together, giving the appearance of friend effect. To alleviate this concern, we control for *LocalCurFrndReviews*, defined as the number of reviews written by current friends in period *t-1* who live in the same city as focal users. If our results are indeed driven by friends going to restaurants together, which is more likely when they are co-located, then *LocalCurFrndReviews* would be significant, and the effect of *CurFrndReviews* will disappear.

ReLogit results are reported in Table 7. The odds ratio for *LocalCurFrndReviews* is insignificant, and that for *CurFrndReviews* is significant and similar to our main analysis, suggesting that the observed effect is unlikely driven by friends going to restaurants together.

Table 1.7. Robustness Check – Going to Restaurants Together

Independent Variables	Odds Ratio
	(SE)
LocalCurFrndReviews $_{i,j,t-1}$	1.070
	(0.180)
CurFrndReviews $_{i,j,t-1}$	2.884***
	(0.311)

NewReviews $_{i,j,t-1}$	1.093***
	(0.004)
FutFrndReviews $_{i,j,t-1}$	2.174***
	(0.253)
N	3,663,479

DV = whether user *i* reviews restaurant *j* in period *t* (*Review$_{ijt}$*). The values in parentheses are standard errors. Control variables were the same as the main analysis but omitted for brevity. + p<0.10, * p<0.05, ** p<0.01, *** p<0.001.

5.4.4. Do Less Popular Restaurants Benefit from Friend Contribution?

Having confirmed the main effect of friend contribution on review quantity, we were also interested in whether the effect of friend contribution differs along restaurant characteristics, in particular their popularity as indicated by the number of reviews received. If the effect is small or none for restaurants that have few reviews, friend effect would not be as useful since new reviews are most needed for restaurants with few existing reviews.

To test the sensitivity of our findings to restaurant popularity, we added an interaction term *CurFrndReviews* * *Log#CumReviews*. ReLogit results are reported in Table 8. We found that the odds ratio for the interaction term is significant and less than 1, suggesting that as the restaurant popularity increases, the effect of friend reviews decreases.

Table 1.8. Further Insights on Review Quantity – Sensitivity to Restaurant Popularity

Independent Variables	Odds Ratio
	(SE)
CurFrndReviews $_{i,j,t-1}$	26.088***
	(8.885)
Log#CumReviews $_{j,t-1}$	1.606***
	(0.014)
CurFrndReviews $_{i,j,t-1}$*Log#CumReviews $_{j,t-1}$	0.639***
	(0.043)
FutFrndReviews $_{i,j,t-1}$	2.234***
	(0.254)
NewReviews $_{j,t-1}$	1.090***
	(0.004)
N	3,663,479

DV = whether user *i* reviews restaurant *j* in period *t* ($Review_{ijt}$). The values in parentheses are standard errors. Control variables were the same as the main analysis but omitted for brevity. + p<0.10, * p<0.05, ** p<0.01, *** p<0.001.

The existence of an interaction effect in a nonlinear model cannot be solely determined by the significance of the interaction term. We followed the common practice to plot the marginal effects of *CurFrndReviews* over *Log#CumReviews* to obtain further evidence. Figure 4 plots the predicted probability of a user reviewing a restaurant when there is a friend review (*CurFrndReviews = 1*) and when there is a stranger review (*CurFrndReviews = 0*), for restaurants with different number of existing reviews *Log#CumReviews*, while holding all other factors at their means. This figure and pair-wise comparison suggest that the effect of a friend review is greater for less popular restaurants, and remains significant up to *Log#CumReviews* = 8 (approximately 3,000 reviews). This suggests that the effect of friend reviews exists across a wide range of restaurant popularity, and is even stronger for restaurants with few reviews.

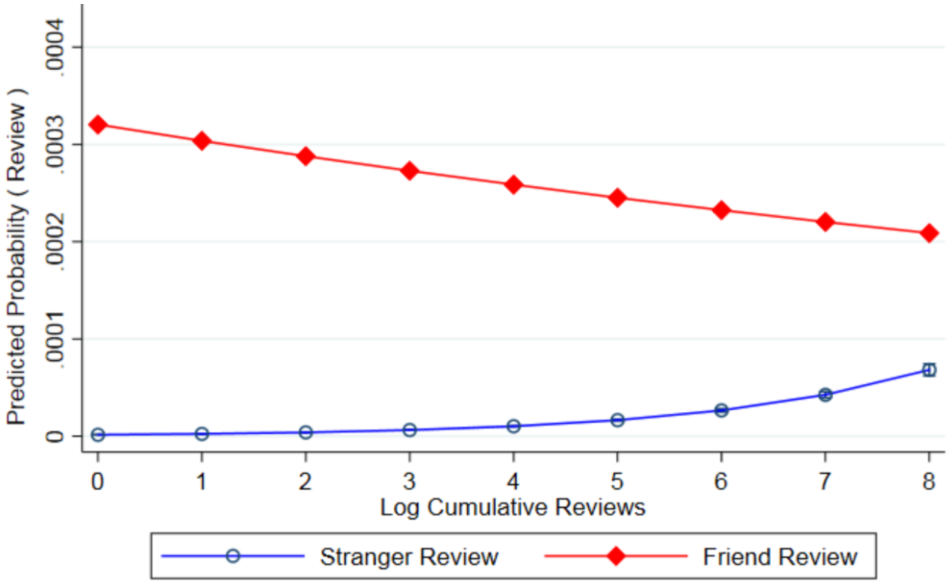


Figure 4.4. Marginal Effect of Current-Friend Reviews as a Function of Cumulative Reviews

6. Analysis on Review Quality

6.1. Friend Effect on Review Quality Using Votes

Review quality is a subjective measure, reflecting a consumer's evaluation of how useful a particular review is in assisting a purchase decision. The literature has predominantly used helpfulness votes received by a review as a proxy for review quality (e.g., Burtch et al. 2017; Otterbacher 2009; Wang et al. 2017). Following the literature, we first used votes as a review quality measure. Yelp has three kinds of votes: *useful*, *funny*, and *cool*. As all three dimensions are related to review quality, we combined them to form *LogVotes*, as a measure for review quality.

We constructed a panel dataset at the user × restaurant level, consisting of users who have offered a review for the restaurants. We used *LogVotes* as the dependent variable, and *CurFrndReviews* and *NewReviews* as independent variables, along with most of the user/restaurant characteristic controls from our analyses of review quantity.

We estimated a panel-OLS model with user fixed effects. The fixed effect controls for the effect of any time-invariant user characteristics. We started with only control variables, then added *CurFrndReviews* and *NewReviews*. Our fixed-effect panel-OLS results are reported in Table 9, Models 1 and 2. As shown, the coefficient for *CurFrndReviews* is positive and significant, suggesting that friend reviews have a greater effect on quality than reviews by strangers. Therefore, ***Hypothesis 2*** is supported.

Table 1.9. Friend Effect on Review Quality – Fixed-Effect OLS

	DV = Log Votes Received *(LogVotes)*		**Robustness Check – Review Length (DV = *LogLength*)**		**Robustness Check - Lagged Friend Reviews**
Independent Variables	**Model 1**	**Model 2**	**Model 3**	**Model 4**	**Model 5**
	Coefficient	Coefficient	Coefficient	Coefficient	Coefficient
	(SE)	(SE)	(SE)	(SE)	(SE)
CurFrndReviews $_{i,j,t-1}$	-	0.115*	-	0.089**	0.093+
	-	(0.049)	-	(0.031)	(0.051)
CurFrndReviews $_{i,j,t-2}$	-	-	-	-	0.099
	-	-	-	-	(0.067)

NewReviews $_{j,t-1}$	-	0.008*	-	0.005+	0.008*
	-	(0.003)	-	(0.003)	(0.003)
#Compliments $_{i,t-1}$	0.043**	0.043**	0.034**	0.033**	0.043**
	(0.016)	(0.016)	(0.012)	(0.012)	(0.016)
#SelfReviews $_{i,t-1}$	0.007*	0.006+	0.011***	0.011***	0.006+
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Log#CumSelfReview $_{i,t-1}$	0.068	0.066	0.180**	0.178**	0.066
	(0.059)	(0.058)	(0.067)	(0.066)	(0.058)
LogTenure $_{i,t-1}$	-0.244	-0.246	-0.415**	-0.417**	-0.247
	(0.189)	(0.189)	(0.136)	(0.136)	(0.190)
Log#Friends $_{i,t-1}$	0.163*	0.166*	0.192**	0.194**	0.167*
	(0.065)	(0.065)	(0.069)	(0.069)	(0.065)
AvgRatingRestaurant $_{j,t-1}$	0.043***	0.043***	0.025*	0.025*	0.043***
	(0.013)	(0.013)	(0.011)	(0.011)	(0.013)
AvgVariRestaurant $_{j,t-1}$	-0.033	-0.025	0.063*	0.068*	-0.025
	(0.034)	(0.034)	(0.028)	(0.027)	(0.034)
Log#CumReviews $_{j,t-1}$	-0.035***	-0.041***	0.001	-0.002	-0.041***
	(0.006)	(0.006)	(0.005)	(0.005)	(0.006)
Promoted $_{j,t-1}$	0.193+	0.179+	0.081	0.072	0.180+
	(0.109)	(0.107)	(0.082)	(0.083)	(0.107)
Claimed $_{j,t-1}$	0.019	0.020	0.022+	0.023+	0.020
	(0.015)	(0.015)	(0.012)	(0.012)	(0.015)
Price $_{j}$	0.093***	0.093***	0.186***	0.186***	0.093***
	(0.014)	(0.014)	(0.011)	(0.011)	(0.014)
Constant	1.576	1.594	7.140***	7.152***	1.598
	(1.252)	(1.256)	(0.896)	(0.896)	(1.259)
Month dummies	included	included	included	included	included
Restaurant category dummies	included	included	included	included	included
Log-Likelihood	-17,903.63	-17,892.86	-14751.65	-14745.08	-17,890.98
Pseudo R-squared	0.458	0.459	0.542	0.543	0.459
N	18,387	18,387	18,387	18,387	18,387

The subscript *t* in the variable names indicates the period in which the review event occurred. For example, *CurFrndReviews*$_{i,j,t-1}$ is the number of current-friend reviews during the period before the focal user's review. DV in Model 5 is log votes received by user *i*'s review on restaurant *j* (*LogVotes*). The values in parentheses are standard errors. + $p<0.10$, * $p<0.05$, ** $p<0.01$, *** $p<0.001$.

6.2. Friend Effect on Review Quality Using Turker Evaluations

Votes can be affected by extraneous factors unrelated to review quality, such as the order in which reviews are displayed or the social relations between voters and the reviewer. Thus, we complemented votes with review quality ratings by participants on Amazon Mechanical Turk (AMT), called “Turkers”. In this AMT study, Turkers rated the reviews from our Yelp data in isolation, thus eliminating much of the extraneous influences on review quality. The details of the study are provided in the Online Appendix.

We run an OLS model with user fixed effects to control for user heterogeneities that could impact review quality. Our fixed-effect OLS results are reported in Table 10. Our results show that the coefficient of *IsFrndReview* is positive and significant, suggesting that friend reviews resulted in higher-quality reviews when compared with stranger reviews. Therefore ***Hypothesis 2*** is again supported.

Table 1.10. Friend Effect on Review Quality (Turker Evaluations) – Fixed-Effect OLS

Independent Variables	Coefficient
	(SE)
IsFrndReviews	0.186*
	(0.085)
Constant	3.350***
	(0.058)
Log-Likelihood	-1147.31
Pseudo R-squared	0.128
N	754

DV= user *i*’ review quality on restaurant *j* (*Quality*). The values in parentheses are standard errors. + p<0.10, * p<0.05, ** p<0.01, *** p<0.001.

6.3. Robustness Checks and Further Insights

6.3.1. Does Friend Contribution Result in Different Review Lengths?

Review length indicates the reviewer’s effort, which can affect the review’s perceived quality. Past research has used the textual length of reviews as another proxy for review quality (Burtch et al. 2018; Pan and Zhang 2011). If friend contribution has an effect on review quality, we expect it to affect review lengths also. We used the same dataset and model as our main analysis on votes.

Fixed-effect Panel-OLS results are shown in Table 9, Model 3 and 4. Consistent with the vote measure, *CurFrndReviews* has a positive effect, suggesting that friend reviews have a stronger effect on review length than stranger reviews.

6.3.2. Do Friend Reviews Have Lagged Effects?

Similar to our analysis of review quantity, we also checked whether our result was robust when we included a lagged effect of friend reviews. Specifically, we added friend reviews in period *t-2* ($CurFrndReviews_{t-2}$) to our main model. Results of fixed-effect Panel OLS are reported in Table 9, Model 5. We find that *CurFrndReviews* in period *t-2* is not significant, suggesting that review quality is only affected by friend reviews in the last period, not earlier. The coefficient of last-period friend reviews is marginally significant, suggesting that our main result on review quality still holds.

7. Discussion and Implications

Online reviews play a crucial role in consumers' decision making, but voluntary contributions of online reviews are scarce. We investigate how the friend effect, in the form of exposing users to reviews written by their online friends, can be a new way of motivating users to write more and higher-quality reviews. We find support for positive effects of friend contribution. Users are three times more likely to provide a review after a friend has written one on the same store, and this effect cannot be solely explained by homophily. Furthermore, the reviews written after a friend's review tend to be of higher quality. We also demonstrate that the types of friends a user responds to depend on her and her friends' tenure and status on the platform.

7.1. Contributions to the Literature

Our findings make several contributions to the academic literature. First, we identify a positive friend effect on the provision of online reviews. Exposing users to reviews written by their friends not only has a sizable effect on the likelihood of users' contributing, but also increases the quality of their contribution. To our knowledge, this is the only known approach to simultaneously increase the quantity and quality of online reviews.

Second, we contribute to the literature of social influence by studying a specific form of social influence, the effect of friend contribution, in a new online context. Online reviews have the characteristics of both public goods and information goods, and are costly to produce, making it a distinctive context for studying friend influence. Our findings are consistent with the predictions of the competitive altruism theory of voluntary contribution – users compete for status by following friends' altruistic contributions with their own high-quality contributions; moreover, as predicted by the competitive altruism theory, a user selects friends to follow based on her and her friend's tenure and status. We believe our novel theoretical perspective and findings broaden the scope of social influence research and hold implications for the social production of user-generated content.

Third, we contribute to the literature on online communities by discovering how the types of online friends users respond to vary by their status and tenure on the platform. We find that new, non-elite users respond to their elite friends, whereas old non-elite users respond to their non-elite friends. New elites respond to nearly all type of friends whereas old elites respond to only their old non-elite friends. To our knowledge, such differential friend-effect patterns have not been documented in the literature. These new findings open doors for future investigations as to the effects of status and tenure on social influence.

7.2. Managerial Implications

Our findings provide several practical implications for managers of online review platforms. First, we provide them a new tool to boost review contribution – to show users reviews written by their online friends. If the platform does not yet support online social networking, it is beneficial to provide it in order to leverage this tool. We note that this tool has many positive attributes: it increases both quantity and quality of reviews, its effects accumulate as there are more friend reviews, and its effect is stronger for stores with fewer existing reviews. Our findings also suggest that friend review recommendations should be personalized: new users who have not achieved elite status should be shown reviews from their elite friends; old users who have not achieved elite status should be shown reviews from their non-elite friends; elite users who are still new should be shown all kinds of friend reviews; and elite users who have been in the community for a long time should only be shown reviews from non-elite friends from the past.

7.3. Limitations and Future Research

Although we have accounted for homophily and many other control variables, we cannot completely rule out the possibility that some unobserved events may explain both the friend and the focal user's reviews. To this end, a randomized field experiment can further this research. Second, we do not have data on consumption or clicks, which prevent us from determining the exact stage at which the friend effect takes place. Third, we have not explored the review text, which can be a good source for more insights on review quality and nuanced user behaviors. Fourth, readers should use caution in generalizing our results to a different notion of online friends, such as those based more on offline relationships than shared interests (e.g. on Facebook). Finally, we employed a widely used snowballing approach to collect our study sample (e.g., Goes et al. 2014; Ma et al. 2013), which can create biases in our study sample, despite our effort to include nearly 80% of all target users.

7.4. Concluding Remarks

Providing an adequate volume of high-quality online reviews is crucial for consumers, businesses, and online review platforms. This study examines a new approach for boosting online review provision by showing users reviews written by their online friends. We demonstrate that such a “friend effect” approach is effective, and has fewer downsides than other known approaches. Our findings hold important implications for research on online reviews, social influence, and user-generated-content communities.

Chapter 2. The Ebb and Flow of Online Word of Mouth

1. Introduction

Online word of mouth (WOM) in the form of online ratings of products and services (henceforth known as goods) has played an influential role in consumer choices. According to a 2013 report by Dimensional Research (2013), about 67 percent of U.S consumers consult online ratings, and an overwhelming 88 percent of them report that their purchasing decisions were influenced by those ratings. There is extensive evidence that online ratings impact sales in industries such as e-commerce, hospitality, and publishing (Chevalier and Mayzlin 2006; Gu et al. 2012; Luca 2011). Given the enormous influence of online ratings on consumer decisions, a crucial issue is whether online ratings are a reliable source of quality information.

There are at least two main concerns over the reliability of online ratings. First, online ratings are written by a highly self-selected group, indicating a *biased representation*. Indeed, existing research shows extremely positive and negative opinions are more likely reported (Gao et al. 2012; Hu et al. 2009). Such biases produce a systematic deviation in the aggregate online ratings such that consumers could be misled about the true quality of goods. Second, aberrations in online ratings, either caused by idiosyncratic ratings or by purposeful manipulations, may cast a long shadow on future ratings such that new ratings may deviate from true quality for a prolonged period. The latter concern, or a *lack of resistance to rating aberrations*, is highly relevant because individuals and businesses are increasingly tempted to game online ratings to their advantage. Anecdotal reports suggest that business owners sometime bribe customers in exchange for favorable reviews of their businesses, or for detrimental reviews of their competitors (The New York Times 2011), and even resort to markets for fake reviews for such purposes (CNET 2016). If a small number of deviant ratings can sway the overall rating of a good, the long-term viability of an online rating system is threatened.

While there is some research on representational biases in online reviews (e.g., Adomavicius et al. 2013; Gao et al. 2012; Hu et al. 2009), there is not much research on aberration resistance. One way to approach this issue is to examine ebb and flow of online ratings. Specifically, to what extent do the new ratings move in the opposite direction of, or "correct," the prior rating aberrations? If the new ratings are at the long term mean, regardless of prior rating aberrations, we say that there is a "full correction"; if the new ratings deviate from the mean in the same direction of recent rating aberrations, we say there is a "partial correction" or "no correction." Conversely, if the new ratings deviate from the mean in the opposite direction, we say there is an "over correction." Given these multiple possibilities, we focus on the following research questions in this paper: *when there is a positive or negative rating aberration in online ratings, do new ratings tend to correct it? If so, by how much?*

The issue of aberration resistance is important, because in the absence of such resistance, online rating systems can deteriorate and eventually become useless as the effects of aberrations compound. Moreover, lack of aberration resistance can also invite gaming of online rating systems, eroding our trust in them. Despite the importance of this issue, there is surprisingly little research on how real-world rating systems behave under rating aberrations.

We approach this issue in two ways. First, we conduct an econometric analysis on a large panel of restaurant ratings on Yelp, a leading online rating site in the U.S. for local businesses. In this observational study, we estimate the effect of last-month's rating aberrations on the valence of next month's ratings, while controlling for restaurant heterogeneities and a few trend/temporal factors. Second, we complement the observational study with randomized online experiments conducted on Amazon's Mechanical Turk, where we ask subjects to rate jokes with exogenously administrated aberrations in their displayed ratings.

With this research, we intend to make the following contributions. First, we hope to shine a light on the issue of the reliability of online rating systems, particularly on their resistance to rating aberrations. As we have mentioned, existing research has already made inroads into the representational biases in online rating systems (Gao et al. 2012; Hu et al. 2009). While such biases are threatening to the accuracy of online ratings, they are static and users of such rating systems may still have a chance to learn and adapt to such biases. The shadow of rating aberrations is a different issue. If uncorrected, the effects of rating aberrations could compound and render the rating systems unusable. Until now, this issue has rarely been addressed in the literature of online WOM. The only other work that comes close to our research questions is Muchnik et al. (2013), who study a binary voting system in an online forum in a field experiment. They manipulate the existing votes by inserting an up or down vote and observe whether such manipulation has a lasting effect on subsequent votes. We discuss in further detail how their setting and research design are different from ours, and why our study adds value to this burgeoning research area.

Second, we also intend to contribute to the novel issue of whether new ratings should partial-, full-, or over-correct prior rating aberrations, especially on the issue of whether the corrective strengthening is symmetric for positive and negative deviations. In this paper, we outline different predictions from a few theoretical streams, and pit them against each other through our observational study and experiments. We believe that our research design for investigating corrective symmetry is an improvement over the research design by Muchnik et al. (2013). In Muchnik et al. (2013), because there are more positive votes than negative ones, an artificial positive vote is inherently a less significant deviation than an artificial negative vote; in contrast, we compare a positive deviation from the (long-term) mean with a same-sized negative deviation from the mean, which are more comparable by construction.

Third, we also contribute to the online WOM literature by studying rating generation for recurring consumables (e.g. restaurants), which involves repetitive consumption and perpetual offerings, as opposed to one-time consumables (e.g. books), which have a fixed quality and a limited shelf time. As far as we know, all the existing studies of rating generations focus on one-time consumables, including movies (Dellarocas et al. 2010; Dellarocas and Narayan 2006; Lee et al. 2015), books (Godes and Silva 2012; Li and Hitt 2008; Wu and Huberman 2010), and music (Wang et al. 2015).[10] Ebb and flow patterns observed among ratings for one-time consumables are probably different from those for recurring consumables due to the importance of life-cycle patterns to one-time consumables (Li and Hitt 2008). Conversely, resistance to rating aberrations is perhaps more important for recurring consumables, due to their perpetual offerings.

The rest of the paper is as follows: we first present relevant literature, conceptual development, and theoretical background. We then present study designs, results, and concluding remarks.

2. Related Literature

An extensive body of literature investigates consequences of online WOM i.e., how online WOM affects consumer decisions and sales (e.g., Chevalier and Mayzlin 2006; Gu et al. 2012; Lu et al. 2013; Luca 2011; Sun 2012). In comparison, there is a much smaller literature on antecedents of online WOM – i.e. what determines the volume and valence of online WOM (e.g., Dellarocas et al. 2010; Goes et al. 2014; Liu et al. 2014; Wang et al. 2015). Our research belongs to the latter category. Within this category, our work is related to three streams of research, online WOM generation, biases in online WOM, and dynamics of online WOM ratings. We discuss them in turn.

The literature of online WOM generation focuses mostly on factors that influence the volume of online WOM. Prior research has identified several factors that may affect the volume of reviews,

---

[10] While several authors also study online reviews for recurring consumables (Luca 2011), their focus is on the impact of reviews on sales rather than rating generations.

such as store/product characteristics (e.g., Dellarocas et al. 2010; Wang et al. 2015), consumption experiences, (e.g., Dellarocas and Narayan 2006), user characteristics (e.g., Goes et al. 2014; Moe and Schweidel 2012), and popularity of an item (e.g., Dellarocas et al. 2010; Sridhar and Srinivasan 2012). These findings help explain why some goods are reviewed more often than others, and why some users tend to write more reviews. Within this literature, characteristics of existing ratings are also included, such as the number of existing ratings (e.g., Dellarocas et al. 2010; Sridhar and Srinivasan 2012), previous average rating (e.g., Ma et al. 2013), and the variance of existing ratings (e.g., Moe and Schweidel 2012). These factors reveal potential dynamics in the arrival rate of new reviews. Our primary focus is on the valence of new ratings. We also focus on a new factor, rating aberrations. We note that rating aberrations are conceptually different from variance-based metrics. The latter does not capture the direction of deviations or episode-level deviations.

Our work is related to a small literature on biases of online WOM. This literature is concerned with whether the expressed opinions reflect the true underlying valuations. Several types of biases are reported in prior studies: One type of bias regards skewed opinion reporting. Hu et al. (2009) report a J-shaped rating pattern in online WOM and argue that this is because consumers only write reviews when they are very satisfied or very dissatisfied. Similarly, Gao et al. (2012) show that, compared to offline ratings, online ratings are more exaggerated and are more concentrated on high-quality providers. Another type of "bias" concerns how expressed opinions are shaped by exogenous factors. Adomavicius et al., (2013) demonstrate an anchor effect which shows that users give higher ratings when the recommendation they receive carries a higher predicted rating. We add to this small literature by focusing on the effect of recent rating aberrations on the valence of subsequent ratings.

Several studies examine the dynamics of online WOM (e.g., Li and Hitt 2008; Moe and Schweidel 2012; Moe and Trusov 2011; Schweidel and Moe 2014). These studies examine the long-term trend of valence and report a declining trend in average ratings over time in book reviews

(Godes and Silva 2012; Li and Hitt 2008). Researchers attribute the declining trend to reviewer self-selection (early readers were more favorable of the product) and purchase errors (Godes and Silva 2012; Li and Hitt 2008). We note that these existing studies focus on one-time consumables, whose rating dynamics are clearly driven by product life cycles, whereas we focus on recurring consumables. Moreover, existing research focuses on long-term trends, whereas we focus on near-term intertemporal dynamics.

Among the literature of dynamics of online WOM, the work most closely related to ours is that of Muchnik et al. (2013). They study a voting system instead of a rating system. Muchnik et al. (2013) randomly insert an initial up or down vote in a voting system for an online reading forum, where readers can vote an "up" or a "down" on others' comments. They show that an artificial down-vote at the beginning has no effect on the aggregate votes as it is followed by a higher proportion of up votes, but an artificial up vote results in a positively bias in the aggregate vote. Despite similar goals between their study and ours, there are many differences. We study a 5-star rating system rather than a binary voting system. We compare a positive and negative aberration relative to the long term mean, while they compare the effects of an up vote and a down vote, which are inherently different in significance because there are usually more positive votes than negative ones (Gao et al. 2012). Our study context is more complex: for example, in the design of Muchnik et al. (2013), readers do not see aggregate votes before they select a comment, whereas in our context, aggregate ratings are posted for all to see.

3. Conceptual Development and Theoretical Background

In a numerical rating system (say ratings are submitted on a 5-star scale), we define rating aberration as a deviation from the accumulative mean (i.e., long-term mean at the moment). Formally, denote $R_t = \{r_{t1}, r_{t2}, \dots r_{tk}\}$ as the set of new ratings in period *t*, and let $\bar{r}_t = \frac{1}{k}\sum_{l=1}^{k} r_{tl}$ be the mean of all ratings in period *t*. Further denote the long-term mean $\bar{\bar{r}}_t =$ as the mean of all ratings that arrived before period *t*. A positive aberration is defined as $\delta_t^+ = \max(\bar{r}_t - \bar{\bar{r}}_t, 0)$ and a negative aberration is defined as $\delta_t^- = \max(\bar{\bar{r}}_t - \bar{r}_t, 0)$. We define the (relative) valence of ratings in period *t* as $y_t = \bar{r}_t - \bar{\bar{r}}_t$.

We are interested in the relationship between rating aberrations in period t-1, namely δ_{t-1}^+ and δ_{t-1}^-, and the valence in period *t*, namely y_t, as indicated by the coefficients $\beta_1 \; and \; \beta_2$ in the following relationship:

$$y_t = \beta_1 \delta_{t-1}^+ - \beta_2 \delta_{t-1}^- + other\ factors$$

For each coefficient, there are four possible scenarios. Taking the coefficient for the positive aberration for example, we have the following four scenarios (see Figure 1):

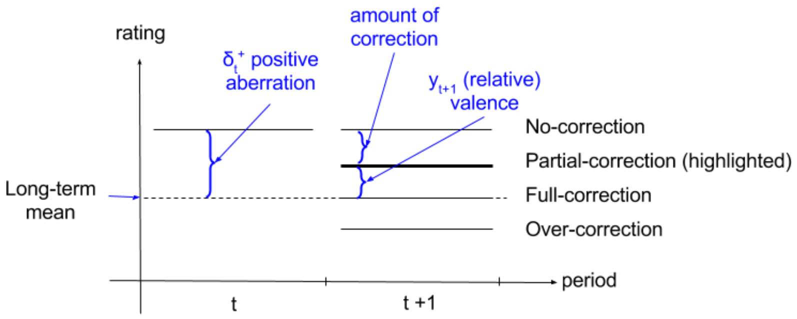


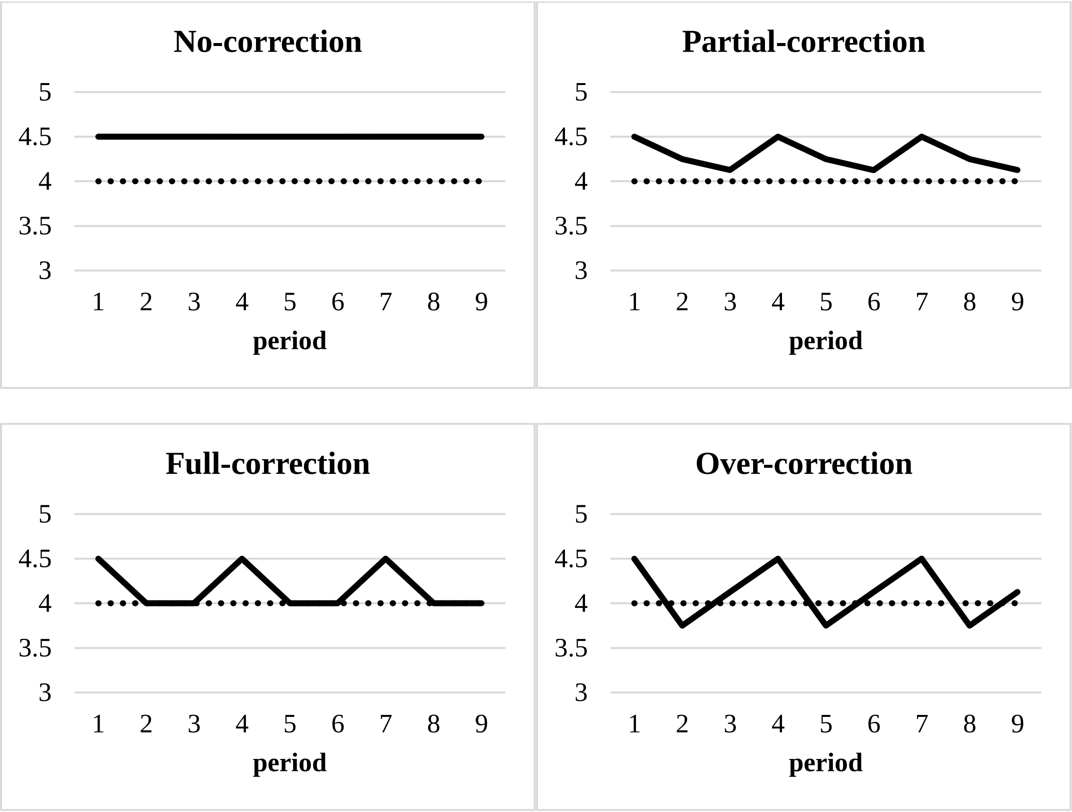
Figure 2.1. Four Types of Correction

1. $\beta_1 = 1$ and $y_{t+1} = \delta_t^+$. The new rating has the same valence as last period's rating. We call this scenario "no-correction" in the sense that the new rating follows the prior rating exactly.
2. $0 < \beta_1 < 1$, and $0 < y_{t+1} < \delta_t^+$. The new rating is lower than the prior rating, but higher than the long-term mean. We call this "partial-correction" because the new rating partially follows the prior rating but adjusts towards the long-term mean.
3. $\beta_1 = 0$ and $y_{t+1} = 0 < \delta_t^+$. The new rating is at the long-term mean. We call this scenario "full-correction" because the new rating is completely unaffected by the prior rating.

4. $\beta_1 < 0$, $y_{t+1} < 0 < \delta_t^+$. The new rating is less than the long-term mean. We call this case “over-correction” because the new rating counters the prior rating with an opposite deviation.

We can also define the *correction percentage* as the amount of correction (relative to the period *t*’s rating) divided by the aberration. The correction percentages for the above four scenarios are 0%, 1%-99%, 100%, and 101%+ respectively.

The four scenarios have different implications for the valence dynamics of online WOM. Figure 2 plots rating dynamics under the four cases, assuming a long-term mean rating of 4 (dashed lines) and a rating aberration of +0.5 every three periods. As seen from this chart, in the no-correction case, new ratings (solid lines in black) never come back to the mean. In the partial-correction case, new ratings slowly come back to the mean. In the full-correction case, new ratings fall right back to the mean. In the over-correction case, new ratings oscillate around the mean. As the correction strength increases, we increasingly see a pattern of “ebb-and-flow.”


No-correction

Partial-correction

Full-correction

Over-correction

Figure 2.2. Rating Dynamics under the Four Correction Scenarios

3.1. The Relationship between Rating Aberrations and Subsequent Rating Valence

We now turn to theories that may potentially explain why rating aberrations may or may not affect the subsequent rating valence. One theory is that new evaluators could also arrive at their conclusions independently without referring to recent ratings. According to this independence theory, we would expect the new ratings to track long-term means rather than recent deviant ratings that is, we would expect a *full-correction*.

An alternative theory is anchoring. The "anchoring effect" is an umbrella term for the phenomenon whereby judgements tend to assimilate around an anchor value provided just before the judgements. Depending on whether the anchor value is related or unrelated to the subsequent judgement, there could be different theoretical explanations. We focus on the former case, as it is more relevant to our context. One explanation is anchoring as *information access*: when an anchor (e.g. a high rating) is provided, information consistent (inconsistent) with the anchor becomes more (less) accessible, which leads to a biased evaluation towards the anchor (Higgins 1996). Another explanation is anchoring as *adjustment* (Tversky and Kahneman 1974). This view holds that when people face decision uncertainty, they use the provided anchor value as a starting point and adjust their evaluation through a search for additional information. This process results in insufficient adjustments due to the search cost, and thus an anchoring effect. A third explanation is that, when anchor values are provided by other individuals, people may view them as social norms, and thus choose to comply with them. A few papers seem to find support for the anchoring effect. For example, Sparks and Browning (2011) show that people rate a store more favorably when positive ratings are presented first. Adomavicius et al. (2013) demonstrate that when a recommender system shows a higher rating, people tend to give higher ratings than if they are shown a lower recommender rating.

The theories for the anchor effect do not concern whether the anchor value deviates from the mean; in fact, they usually focus on cases where there is no obvious norm or preconception. Thus anchor theories in general predict that judgement will center around the anchor value – a *no-correction* case.

When there is a clear sense of mean, the anchor-and-adjustment theory of Tversky and Kahneman (1974) can be extended to provide a directional adjustment prediction. By this theory, when anchor value is above the mean, thereby resulting in a positive aberration, the search process is more likely to yield unfavorable information, thus the direction of adjustment will thus be down. (The argument for a negative aberration is similar and thus omitted). The anchor-and-adjustment theory suggests that the adjustment is often insufficient, and people will thus arrive at a value above the mean on average – *a partial-correction* case.

There are also reasons to suspect that new evaluators may over-correct recent rating aberrations. First, the Expectation Confirmation Theory (ECT) may predict an over-correction (Anderson and Sullivan 1993). According to ECT, the post-evaluation on goods performance is determined by the extent to which goods performance exceeds expectation (Oliver 1980). Positive confirmation occurs when the perceived performance exceeds the expectation, and vice versa (Anderson and Sullivan 1993). In our study context, when there is a positive aberration, evaluators see high recent ratings and form an unrealistically high expectation. This high expectation is most likely disconfirmed by a less-than-stellar performance, which results in disappointment and a lower rating than they would have given without such high expectations. In a similar vein, a negative aberration would result in pleasant surprises and excessively high ratings – *an over-correction case*.

Over-correction may also be explained by customer loyalty. A loyal customer may be motivated to protect a store's reputation and counter negative ratings with excessively high ratings, to offset the damage done by recent negative ratings. This is especially true when many people only look at aggregate ratings. Such vigilant over-correction behaviors can be carried out by loyal customers spontaneously or they can be prompted by store owners, or even elicited through “bribery.”

In sum, given the countervailing theoretical arguments regarding the effect of rating aberrations, we formulate our first research question as:

Q1: *When there is a positive or negative rating aberration in online ratings, do new ratings tend to correct it? If yes, what is the strength of the correction (is it a partial-, full-, or over-correction)?*

3.2. Symmetry of Corrections

Provided that there is some level of correction, the next question is whether the corrections for the positive and negative aberrations are symmetric. This is important because it relates to whether there will be systematic accumulation of aberrations in online ratings. Prior studies (e.g., Dellarocas and Wood 2008; Hu et al. 2009) have already shown a positivity bias in online WOM – that online ratings tend to be higher than true opinions. Our focus is different. We take the overall positivity bias of online ratings as given, and ask whether deviations from online rating norms trigger equal-strength corrections.

One seemingly plausible argument may be that since evaluators are more likely to report a positive rating than a negative rating, negative aberrations will be more likely corrected than positive ones, as correcting negative ones helps the establishment of a distinctive online identity (e.g., Dellarocas et al. 2010; Zeng and Wei 2013). However, this argument may not apply, because we define aberrations relative to mean ratings, which means that people are as likely to report a rating above the mean as below. Thus, there is no ex ante asymmetry in the likelihood of posting a more or less favorable rating relative to the mean. It also means that a positive aberration and a negative one are roughly comparable. This is in contrast with the study by Muchnik et al. (2013), which looks at the effect of a one-up vote versus a one-down vote. In their case, an up vote has a much higher probability of natural occurrence, which predetermines that a down vote is not only more noticeable, but also more likely to be corrected.

A potential reason for asymmetric correction is the existence of entities that are only interested in seeing high (or low) ratings. Loyal customers are only interested in high ratings. Their abundance will lead to a much stronger correction for negative aberrations. Conversely, loyal customers of competitors are only interested in correcting positive aberrations. In sum, the correction for a negative aberration is not necessarily stronger than that for a positive aberration.

Q2: *Is the size of correction for a negative aberration the same as that for a positive aberration? If not, which effect is stronger?*

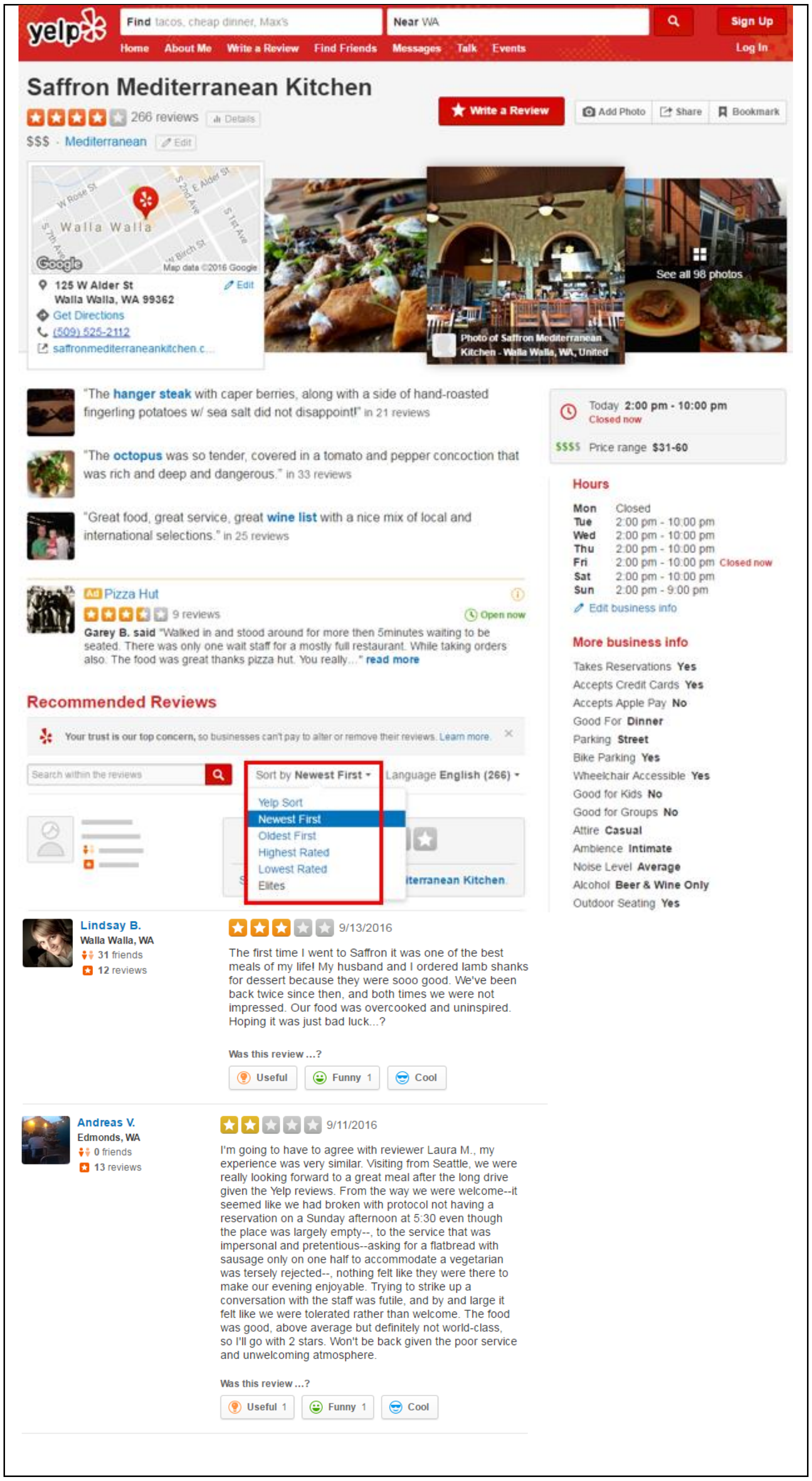
4. Econometric Analysis Using Yelp Data

4.1. Empirical Context and Data

We begin by analyzing a large field dataset, which we collect from Yelp, one of the largest and most successful online review sites in the world. As of 2016, the site has 135 million monthly visitors and 95 million reviews. Users can friend, follow, and send compliments to one another. They may also vote on reviews written by others. Each year, the site selects elite reviewers based on the quality and quantity of the previous year's reviews, and the selected users are honored with an elite badge. Yelp ratings are contributed by volunteers who are typically patrons of stores. The average number of reviews per reviewer account at Yelp is about 25, which is much higher than that at Citysearch, where the average is 2 (Wang 2010). The friend networking also stimulates review production – 0.62% friend reviews are followed up by a user's own review, resulting in 3,567 reviews by 9,766 users in about 7 months (Ke and Liu 2015).

For each store, Yelp publishes an overall average rating in stars in increments of 0.5 stars, along with a list of ratings and corresponding reviews. Yelp offers several options for users to sort the ratings and reviews, including *Newest First*, *Oldest First*, *Highest Rated*, *Lowest Rated*, *Elites*, *and*

Yelp Sort. The default sorting is *Yelp Sort* which is based on recency, user voting, and other review quality factors (Yelp Support Center). While the exact sorting criteria of *Yelp Sort* is unknown, anecdotal observation of *Yelp Sort* results suggests that recency is a primary consideration in *Yelp Sort*. Figure 3 illustrates a typical store page on Yelp.



Saffron Mediterranean Kitchen

266 reviews

$$$ · Mediterranean

125 W Alder St
Walla Walla, WA 99362
Get Directions
(509) 525-2112
saffronmediterraneankitchen.c...

"The **hanger steak** with caper berries, along with a side of hand-roasted fingerling potatoes w/ sea salt did not disappoint!" in 21 reviews

"The **octopus** was so tender, covered in a tomato and pepper concoction that was rich and deep and dangerous." in 33 reviews

"Great food, great service, great **wine list** with a nice mix of local and international selections." in 25 reviews

Ad Pizza Hut
9 reviews
Open now
Garey B. said "Walked in and stood around for more then 5minutes waiting to be seated. There was only one wait staff for a mostly full restaurant. While taking orders also. The food was great thanks pizza hut. You really..." **read more**

Today **2:00 pm - 10:00 pm**
Closed now

$$$$ Price range **$31-60**

Hours

Mon	Closed
Tue	2:00 pm - 10:00 pm
Wed	2:00 pm - 10:00 pm
Thu	2:00 pm - 10:00 pm
Fri	2:00 pm - 10:00 pm Closed now
Sat	2:00 pm - 10:00 pm
Sun	2:00 pm - 9:00 pm

Edit business info

More business info

Takes Reservations	**Yes**
Accepts Credit Cards	**Yes**
Accepts Apple Pay	**No**
Good For	**Dinner**
Parking	**Street**
Bike Parking	**Yes**
Wheelchair Accessible	**Yes**
Good for Kids	**No**
Good for Groups	**No**
Attire	**Casual**
Ambience	**Intimate**
Noise Level	**Average**
Alcohol	**Beer & Wine Only**
Outdoor Seating	**Yes**

Recommended Reviews

Your trust is our top concern, so businesses can't pay to alter or remove their reviews. Learn more.

Lindsay B.
Walla Walla, WA
31 friends
12 reviews

9/13/2016

The first time I went to Saffron it was one of the best meals of my life! My husband and I ordered lamb shanks for dessert because they were sooo good. We've been back twice since then, and both times we were not impressed. Our food was overcooked and uninspired. Hoping it was just bad luck...?

Was this review ...?

Andreas V.
Edmonds, WA
0 friends
13 reviews

9/11/2016

I'm going to have to agree with reviewer Laura M., my experience was very similar. Visiting from Seattle, we were really looking forward to a great meal after the long drive given the Yelp reviews. From the way we were welcome--it seemed like we had broken with protocol not having a reservation on a Sunday afternoon at 5:30 even though the place was largely empty--, to the service that was impersonal and pretentious--asking for a flatbread with sausage only on one half to accommodate a vegetarian was tersely rejected--, nothing felt like they were there to make our evening enjoyable. Trying to strike up a conversation with the staff was futile, and by and large it felt like we were tolerated rather than welcome. The food was good, above average but definitely not world-class, so I'll go with 2 stars. Won't be back given the poor service and unwelcoming atmosphere.

Was this review ...?

Figure 2.3. A Typical Store Page on Yelp

We collected restaurant reviews in the state of Washington (WA) on a periodic basis between January 2012 and April 2014. We searched "WA" and "restaurant" to get the list of all restaurants in Washington state, which resulted in a total of 8,943 restaurants. We also collected detailed information on each restaurant, including monthly aggregate ratings and the total number of reviews; detailed information of each review (e.g., review content, rating given, when posted, by whom, and votes and compliments received); and restaurant information (e.g., price range, location, and category). We excluded restaurants which had no new ratings for two successive months. As a result, 6,519 restaurants remained for our analysis, with about 2,488 restaurants in each period (1 month) and a total of 64,684 restaurant-months.

4.2. Model and Variables

We are interested in the effect of positive and negative aberrations on (relative) valence of subsequent ratings. To address this issue, we construct a panel dataset at the restaurant-month level for restaurants in Washington state between January 2012 and April 2014. Because our dependent variable is continuous, we estimate a panel data OLS model with restaurant fixed effects. The fixed effects allow us to control for unobserved time-invariant characteristics of restaurants such as ambience.

Formally, let $Valence_{it}$ denote the mean of new ratings for restaurant *i* in period *t* minus the mean of all ratings for restaurant *i* up to period *t-1*. We define $Aberration_{i,t-1}$ as the mean of ratings for restaurant *i* in period *t-1* minus the mean of all ratings for restaurant *i* up to period *t-2*. We define $PosAberration_{i,t-1} = \max(Aberration_{i,t-1}, 0)$ and $NegAberration_{i,t-1} = \max(-Aberration_{i,t-1}, 0)$, that is, $PosAberration_{i,t-1}$ captures the positive deviation from the mean whereas $NegAberration_{i,t-1}$, the negative deviation. We model the valence for restaurant *i* in period *t* as:

$$Valence_{it} = \alpha_i + \beta_1 PosAberration_{i,t-1} + \beta_2 NegAberration_{i,t-1} + \beta_3 Controls_{i,t-1} + \varepsilon_{it} \quad (1)$$

where α_i is a fixed restaurant-specific constant, $Controls_{i,t-1}$ consists of lagged control variables, and ε_{it} is a normally distributed error term.

We note that by using the relative rather than the absolute valence as the dependent variable, we avoid capturing the effect of "regression to the mean" – which states that if we get an extreme value on one observation, the next one tends to move towards the mean. If ratings were purely random, "regression to the mean" would imply that a new rating would have an expected value *at the mean*, so that aberrations would not have a significant effect on the relative valence. By using relative valence as a dependent variable, we hope to capture effects beyond "regression to the mean."

We include several time-variant control variables, including the number of last period ratings. We use Yelp age (defined as number of years since the restaurant joined Yelp) to control for restaurant lifecycle. To control for system-wide temporal shocks to ratings, we include month dummies. Table 1 provides descriptions and summary statistics of variables used in our models. Prior to estimating the models, we conducted collinearity tests and found no signs of collinearity (VIF < 3).

Table 2.1. Variable Definitions and Descriptive Statistics – Observational Study

Variables	**Definition**	**Mean**	**Std. Deviation**	**Min**	**Max**	**# Obs**
Valence	Mean of new rating in period *t* minus mean of all ratings up to period *t-1* (i.e., long-term mean)	-0.044	1.000	-4	4	64,684
Lag positive aberration	Max($Aberration_{i,t-1}$, 0), where $Aberration_{i,t-1}$ is the mean of new ratings in period *t-1* minus the mean of all ratings up to period *t-2*	0.359	0.526	0	4	64,684
Lag negative aberration	Max (-$Aberration_{i,t-1}$, 0)	0.393	0.663	0	4	64,684
# of last period ratings	# of new ratings received in period *t-1*	3.022	3.485	1	117	64,684
Log # of all ratings	Log (1+ # of all ratings up to period *t-1*)	4.136	1.048	1	8	64,684

Accumulative rating	Mean of all ratings up to period *t-1*	3.669	0.579	1	5	64,684
Yelp age	# of years restaurant resided on Yelp up to period *t-1*	3.102	2.322	0	9	64,684

We omitted the statistics of 25 month dummies for brevity.

4.3. Results

We run two versions of the OLS model: a basic version with just controls, and a full version. Our results are reported in Table 2. As seen from this table, the coefficient for the positive aberration is negative and significant (*b*= -0.028, *p*<0.01), thus a positive aberration of 1 is followed by new ratings that are 0.028 below the mean on average. The coefficient for negative aberration is significant and positive (*b*=0.021, *p*<0.01); thus a negative rating aberration of 1 is followed by new ratings that are, on average, 0.021 above the mean. In sum, an over-correction is observed after both positive and negative aberrations. Though the size of the overcorrection is small relative to the aberration, the fact that new ratings move to the opposite side of the mean suggests that the rating system will revert quickly to the mean after the aberration (also see Figure 2).

Table 2.2. Fixed-effect OLS Regression on Valence – Observational Study

Independent Variables	**Model 1**	**Model 2**
	Coefficient (SE)	**Coefficient (SE)**
Lag positive aberration		-0.028** (0.011)
Lag negative aberration		0.021** (0.008)
# of last period ratings	-0.001 (0.001)	-0.001 (0.001)
Log # of all ratings	-0.092** (0.031)	-0.098** (0.032)
Accumulative rating	-1.155*** (0.026)	-1.153*** (0.026)
Yelp age	0.033+ (0.019)	0.034+ (0.019)
Constant	4.478*** (0.134)	4.495*** (0.136)
Log-likelihood	-82,817.77	-82,798.00

Adjust R-squared	0.157	0.158
N	64,684	64,684

The values in parentheses are robust standard errors clustered by restaurants. +p<0.10, * p<0.05, ** p<0.01, *** p<0.001.

We further compare the coefficients for the positive and negative aberrations. An F-test shows that the two coefficients are not statistically different (*F*= 0.23, *p*=0.634). This suggests that the strength of corrections after negative and positive aberrations are not significantly different.

4.3.1. Effects of Aberration Size

Evaluators might respond to large and small rating aberrations differently. For example, small aberrations might not be easy to spot, and larger ones might evoke strong emotions. To test the effect of large and small aberrations, we use the same dataset and, based on the distribution of aberration, code aberration into three degrees: small aberration if aberration is less than or equal to 0.5; medium aberration if aberration is greater than 0.5 and less than or equal to 1; and large aberration if aberration is greater than 1.

OLS results are reported in Table 3. We find that the coefficients for small positive and negative aberrations are insignificant, suggesting that small aberrations are not corrected. The coefficients for both medium and large positive aberrations are negative and significant (*b*= -0.029, *p*< 0.10; *b*= -0.032, *p*< 0.05 respectively). A larger positive aberration of 1 is followed by smaller new ratings than a medium positive aberration of 1, demonstrating stronger correction. Only a large negative aberration is corrected, as the coefficient for it is positive and significant (*b*= 0.023, *p*< 0.01).

Table 2.3. Fixed-effect OLS Regression on Valence – Aberration Size

Independent Variables	**Model 1**	**Model 2**
	Coefficient (SE)	**Coefficient (SE)**
Lag small positive aberration		-0.036 (0.029)

Lag medium positive aberration		-0.029+ (0.015)
Lag large positive aberration		-0.032* (0.013)
Lag small negative aberration		0.019 (0.031)
Lag medium negative aberration		-0.003 (0.017)
Lag large negative aberration		0.023** (0.009)
# of last period ratings	-0.001 (0.001)	-0.001 (0.001)
Log # of all ratings	-0.092** (0.031)	-0.099** (0.032)
Accumulative rating	-1.155*** (0.026)	-1.153*** (0.026)
Yelp age	0.033+ (0.019)	0.034+ (0.019)
Constant	4.478*** (0.134)	4.499*** (0.137)
Log-likelihood	-82,817.77	-82,796.14
Adjust R-squared	0.157	0.158
N	64,684	64,684

We estimate a fixed-effect model on valence. The values in parentheses are robust standard errors clustered by restaurant. + $p<0.10$, * $p<0.05$, ** $p<0.01$, *** $p<0.001$.

4.4. Robustness Tests

4.4.1. Falsification Test

One potential concern about our findings is that they may be merely a statistical phenomenon, as a natural random process also displays a tendency of "regression to the mean." We have specified our model in a way that will not capture such spurious findings. To further address this concern, we conduct a falsification test as follows. We use the same dataset but randomly shuffle dependable variables (i.e., valence). This effectively creates a random pairing of explanatory variables and dependable variables. If our findings are purely driven by a statistical "regression to the mean," we would expect to see the same results after the perturbation. Otherwise, our findings suggest a true underlying dynamic pattern.

OLS results using the shuffled dataset are reported in Table 4. We find that the coefficients for both positive and negative aberrations are insignificant, rejecting the belief that our findings are merely a statistical phenomenon.

Table 2.4. Robustness Tests – Falsification Test

Independent Variables	**Model 1**	**Model 2**
	Coefficient (SE)	**Coefficient (SE)**
Lag positive aberration		0.009 (0.011)
Lag negative aberration		-0.006 (0.008)
# of last period ratings	-0.001 (0.001)	-0.001 (0.002)
Log # of all ratings	-0.001 (0.031)	0.001 (0.031)
Accumulative rating	-0.039 (0.027)	-0.039 (0.027)
Yelp age	-0.015 (0.019)	-0.016 (0.019)
Constant	0.169 (0.138)	0.161 (0.140)
Log-likelihood	-80,609.12	-80,607.28
Adjust R-squared	0.17135	0.17137
N	63,325	63,325

We estimate a fixed-effect model on valence. The values in parentheses are robust standard errors clustered by restaurant.
+ p<0.10, * p<0.05, ** p<0.01, *** p<0.001.

4.4.2. Alternative Measurement of Aberrations

Ideally, the ratings used to calculate aberrations are what evaluators see. When a period has no new ratings, Yelp will show older ratings and evaluators are likely to base their decisions on these older ratings. In our main analysis, we have dropped a restaurant-month observation whenever there are no new ratings in period *t-1*. Here we test an alternative measurement of aberration where we use *t-2* aberrations to substitute *t-1* aberrations when *t-1* aberrations are missing.

We used the same model specification as the main analysis except we use the alternative measure for aberrations. Our results are reported in Table 5. According to this table, the coefficients for both positive and negative aberrations remain significant (*b*= -0.028, *b*= 0.032 respectively). An F-test suggests that the coefficients for positive and negative aberrations are not statistically different (*F*= 0.07, *p*= 0.788) and qualitatively similar to our main findings. Hence, our results are robust to this alternative measurement of aberrations.

Table 2.5. Robustness Test – Measure Aberrations Differently

Independent Variables	**Model 1**	**Model 2**
	Coefficient (SE)	**Coefficient (SE)**
Lag positive aberration		-0.028** (0.009)
Lag negative aberration		0.032*** (0.007)
# of last period ratings	-0.001 (0.001)	-0.001 (0.001)
Log # of all ratings	-0.097*** (0.029)	-0.105*** (0.030)
Accumulative rating	-1.205*** (0.023)	-1.198*** (0.023)
Yelp age	0.028 (0.018)	0.030+ (0.018)
Constant	4.662*** (0.121)	4.658*** (0.122)
Log-likelihood	-104,600.8	-104,562.2
Adjust R-squared	0.153	0.154
N	79,126	79,126

We estimate a fixed-effect model on valence. The values in parentheses are robust standard errors clustered by restaurant.
+ p<0.10, * p<0.05, ** p<0.01, *** p<0.001.

5. Experimental Studies Using Joke Ratings

One of the challenges of using field data is that we cannot be completely sure that evaluators notice the aberrations and make purposeful corrections as a result. In other words, there could still be unknown dynamics that gives a false impression that the evaluators are making corrections (e.g., imagine a process that draws some favorable evaluators in one month but unfavorable ones in the next, or weather patterns that cause mood swings among evaluators from one period to another).

To ease such concerns, we complement the field data study with two experimental studies, an original experiment (Experiment 1) and a follow-up one (Experiment 2), where we administrate randomized exogenous rating aberrations and study their impact on subjects' rating behaviors.

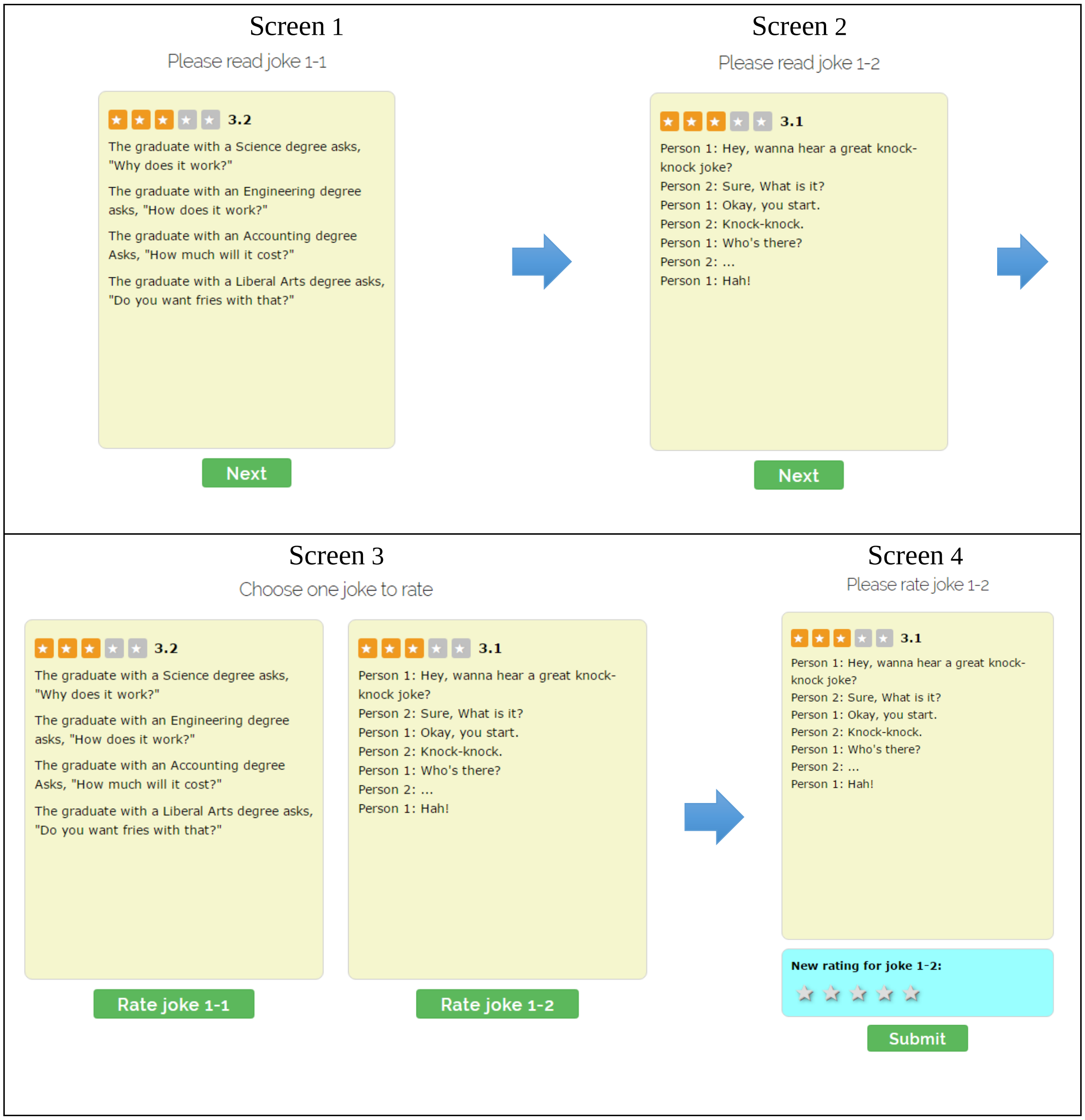
5.1. Experiment 1 (The Original Experiment)

5.1.1. Experiment Design

There are several challenges associated with replicating the context of field studies in experimental settings. In order to fit the rating task in an experiment session, many common tasks such as rating restaurants, professional services, books, and movies cannot be used. Furthermore, we must also have a repository of existing ratings so that we can establish baseline mean ratings and administrate aberrations. We choose rating jokes as our experiment task because they are short enough to do in an experiment. Moreover, we find a set of jokes from Jester Online Joke Recommendation System database (http://eigentaste.berkeley.edu/dataset), which have been extensively rated for recommender system research. We use their dataset 2+ of 150 jokes with over 500,000 new ratings from 79,681 total users between November 2006 and November 2012. Among 150 jokes, 22 were never displayed or rated; the average number of ratings per joke received is over 3,900 and the average ratings range from 1.8 to 3.4, with a mean standard deviation of 5.03.

Another challenge we faced was the rare-event nature of online ratings. On Yelp, restaurant ratings typically follow a visit to the restaurant, but the probability of a user providing a rating after visiting a restaurant is very low. If we were to replicate such rare behaviors faithfully, we would need a very large sample to obtain a reasonable number of ratings. To ensure an adequate number of ratings, we use a discrete choice framework where we ask subjects to choose one joke from each pair to offer a rating. This way we preserve some degree of freedom in choosing, while ensuring the number of joke ratings we can use.

Our design of the experiment task is as follows. We present each subject 16 pairs of jokes, one pair at a time. For each pair, we ask subjects to choose one joke to offer a new rating on a scale of one to five stars. Subjects first read the two jokes one by one (screens 1 and 2). They then choose one of the two jokes (screen 3) to enter a new rating (screen 4). At the end of all 16 pairs, subjects answer several manipulation-check and background questions.



Screen 1

Please read joke 1-1

3.2

The graduate with a Science degree asks, "Why does it work?"

The graduate with an Engineering degree asks, "How does it work?"

The graduate with an Accounting degree Asks, "How much will it cost?"

The graduate with a Liberal Arts degree asks, "Do you want fries with that?"

Screen 2

Please read joke 1-2

3.1

Person 1: Hey, wanna hear a great knock-knock joke?
Person 2: Sure, What is it?
Person 1: Okay, you start.
Person 2: Knock-knock.
Person 1: Who's there?
Person 2: ...
Person 1: Hah!

Screen 3

Choose one joke to rate

3.2

The graduate with a Science degree asks, "Why does it work?"

The graduate with an Engineering degree asks, "How does it work?"

The graduate with an Accounting degree Asks, "How much will it cost?"

The graduate with a Liberal Arts degree asks, "Do you want fries with that?"

3.1

Person 1: Hey, wanna hear a great knock-knock joke?
Person 2: Sure, What is it?
Person 1: Okay, you start.
Person 2: Knock-knock.
Person 1: Who's there?
Person 2: ...
Person 1: Hah!

Screen 4

Please rate joke 1-2

3.1

Person 1: Hey, wanna hear a great knock-knock joke?
Person 2: Sure, What is it?
Person 1: Okay, you start.
Person 2: Knock-knock.
Person 1: Who's there?
Person 2: ...
Person 1: Hah!

New rating for joke 1-2:

Figure 2.4. Screens of Rating Jokes

As with real-world online rating systems, the jokes are displayed with a rating. This provides a chance for us to introduce some random rating aberrations. We adopt a between-subject design. For each pair of jokes, half of the subjects are randomly assigned to the treatment group, where joke 2 (the "focal joke") is displayed with a manipulated rating (i.e., an aberration). The other half is assigned to the control group, where the mean rating from the database (which we call the "true" rating) is shown. We always display the true rating for joke 1. To avoid the effect of display order, we also randomize the order of the two jokes so that half of the subjects will see joke 1 first (or the left joke when the two are shown together), and the other half will see joke 2 first.

The effect of an aberration may be different depending on whether the aberration is positive or negative, and on whether true ratings are both high (HH), both low (LL), high-low (HL), or low-high (LH). We therefore use 4 pairs of jokes for each true rating pattern (HH, LL, HL or LH). We manipulate joke 2 in the feasible direction, that is, the manipulated rating patterns will be HL, LH, HH, and LL respectively. In sum, we have a 4 (HH, LL, HL, and LH) x 2 (with and without aberration) x 2 (left / right) between-subject design. We ask each subject to go through 16 pairs of jokes (thus all 16 conditions, albeit with different jokes), with the order of pairs randomized. By comparing the ratings given to the focal joke in treatment and control groups of the same joke pair, we can obtain the effect of rating aberrations on rating valence.

Based on the distribution of joke ratings, we categorize jokes into H and L rating groups, with average ratings ranging from 3.2 to 3.4 and 2 to 2.6 respectively. We set the size of aberrations to either +1 (up treatment) or -1 (down treatment), so that a joke with a true rating of 3.3 and a down treatment will be displayed with a rating of 2.3 in the treatment group (we always show numeric ratings next to stars). We choose 16 pairs of jokes out of 120 jokes available (8 jokes were included in the "gauge set" and 22 jokes were never displayed and rated, so those 30 jokes were suggested for removal at the Jester website. See Goldberg et al. 2001), so that we have 4 pairs for each true rating pattern and the jokes in each pair have roughly similar lengths and can be fit on one screen.

We conduct our Experiment 1 on Amazon Mechanical Turk, a global crowdsourcing labor market where workers (called Turkers) get paid for doing micro online tasks such as filling out a survey or classifying an image. To ensure the quality of responses, only U.S. Turkers with over 95% task acceptance rates and at least 50 completed assignments were allowed to participate in this study. They are paid 50 cents for rating 16 pairs of jokes, which translates into an average hourly wage of $2.5, which is on par with other Turk assignments. A total of 603 people completed the study within 4 days. We used manipulation checks and time stamps to exclude those who reportedly did not notice the ratings of jokes (86 subjects). Five hundred and seventeen subjects remain usable for our analysis. Among included subjects, 48% are female. The summary statistics are reported in Table 6.

Table 2.6. Variable Definitions and Descriptive Statistics – Experiment 1

Variables	**Definition**	**Mean**	**Std. Deviation**	**Min**	**Max**	**# Obs**
Valence	Actual rating given to focal joke minus true rating	0.190	1.146	-2.4	3	3,671
UpTreatment	Whether focal joke is manipulated up	0.217	0.413	0	1	3,671
DownTreatment	Whether focal joke is manipulated down	0.281	0.449	0	1	3,671
GenderFemale	Subject gender: 1=female; 0=male	0.453	0.498	0	1	3,671
AgeGroup	Which age group subject belongs to: Group 1=18-24 year old; Group 5=65-74 year old	2.727	1.105	1	6	3,671
SkipSomeJokes	Whether subject skips all jokes: 1 yes, otherwise 0	0.1	0.3	0	1	3,671
UpTreatment * SkipSomeJokes	Interaction between UpTreatment and SkipSomeJoke	0.025	0.156	0	1	3,671
DownTreatment * SkipSomeJokes	Interaction between DownTreatment and SkipSomeJoke	0.024	0.152	0	1	3,671

5.1.2. Models and Results

We are interested in the effect of up or down treatments (aberrations) on the actual rating valence. Let $Valence_{ik}$ denote the actual rating given by subject *I* for the focal joke of joke-pair *k*, minus the true rating, and $UpTreatment_{ik}$ ($DownTreatment_{ik}$) denote whether the focal joke's rating is up (down) treated. If the subject is assigned to a control group, then both $UpTreatment_{ik}$ and $DownTreatment_{ik}$ are 0. Otherwise, one of them will be 1, depending on whether the treatment is up or down.

Because the dependent variable is continuous, we estimate a standard OLS regression with joke-pair fixed effects. The fixed effects allow us to control for unobservable characteristics of the two jokes in the pair. Formally, our model is:

$$Valence_{ik} = \alpha_k + \beta_1 UpTreatment_{ik} + \beta_2 DownTreatment_{ik} + \beta_3 Controls_{ik} + \varepsilon_{it} \quad (2)$$

In equation (2), $Controls_{ik}$ denotes a set of control variables, including the subject's gender, age group, and whether he/she self-reported having skipped at least some jokes (*SkipSomeJokes*). Table 6 provides a description and the summary statistics of the variables used in our models.

Based on equation (2), we estimate three models: Model 1 includes just treatment variables, Model 2 adds control variables, and Model 3 adds interaction terms: the interaction between *SkipSomeJokes* and treatment indicators. Table 7 reports the results of the three models. Because the results are highly consistent, we interpret Model 3 as an example. Our results show that the coefficient of *UpTreatment* is positive and significant across three models. In particular, an up treatment by 1 point is followed by ratings that are, on average, 0.358 higher than the true rating and 0.642 lower than the manipulated rating. Thus, we have a case of a *partial-correction*. The coefficient for $DownTreatment$ is negative and significant. In particular, a negative aberration is followed by ratings that are, on average, 0.238 lower than the true rating, but 0.762 higher than the displayed average rating (= true rating – 1). Overall, we find a partial-correction for both positive and negative aberrations. We note that the interaction between *DownTreatment* and *SkipSomeJokes* is negative and marginally significant (*b*= -0.274, *p*= 0.075), suggesting that when an evaluator is paying attention, he or she will rate the joke closer to the mean when the focal joke is down treated. This is consistent with the intuition that correction strength will be stronger if evaluators take the rating task more seriously. We note that the interaction with positive aberrations is not significant but in the right direction. This is possibly because positive aberrations are less discernable.

Table 2.7. Fixed-effect OLS Regression on Valence – Experiment 1

Independent Variables	**Model 1**	**Model 2**	**Model 3**
	Coefficient (SE)	**Coefficient (SE)**	**Coefficient (SE)**
UpTreatment	0.370*** (0.059)	0.371*** (0.059)	0.358*** (0.062)
DownTreatment	-0.262*** (0.048)	-0.261*** (0.048)	-0.238*** (0.050)
GenderFemale		-0.063+ (0.038)	-0.065+ (0.038)
AgeGroup		-0.003 (0.017)	-0.003 (0.017)
SkipSomeJokes		0.137* (0.062)	0.174* (0.087)
UpTreatment * SkipSomeJokes			0.119 (0.153)
DownTreatment * SkipSomeJokes			-0.274+ (0.154)
Constant	3.134*** (0.026)	3.158*** (0.055)	3.153*** (0.055)
Adjust R-squared	0.143	0.144	0.145
N	3,671	3,671	3,671

+ p<0.10, * p<0.05, ** p<0.01, *** p<0.001.

As with the field data study, we also test the difference in the effects of *UpTreatment* and *DownTreatment*, and our results show that the two effects are not statistically different in either model (*F*= 2.01, *p*= 0. 156; *F*= 2.09, *p*= 0.149; and *F*= 2.24, *p*= 0.135 respectively).

We noticed that while we find evidence of corrections in both the observational study and the original experiment, the sizes of corrections are different. In particular, if expressed in percentage terms (valence), we find that a positive (negative) aberration results in a correction percentage of about 102.8% (103.2) in the observational study, and the corresponding percentages in the experiment is around 63% (74%). While these correction percentages both suggest that new ratings will likely revert to the mean quickly, we find a weaker correction in the experiment. There are some possible reasons. To submit an over-correction rating, the evaluators must have a strong opinion regarding the recent ratings, and are likely to take the online ratings of the store very seriously. On Yelp, evaluators are self-selected, and a lot of evaluators (especially elite reviewers) are members of the online review community and proud of their reviews. They probably have strong feelings toward the reviewed local businesses. We argue that they are more likely to submit a corrective rating than subjects in our experiment, who are compelled to provide a rating and who may therefore not care about the overall ratings of the jokes as much as Yelpers do about the local stores they rate. If our argument is correct, then participants in our experiment may be less keen on submitting a corrective rating. Indeed, consistent with the argument that the more committed evaluators are, the more likely they are to submit a corrective rating, we find through the interaction term (*DownTreatment* * *SkipSomeJokes* in Table 7) that some of the more "*serious and committed*" of our participants show a higher tendency to correct than their less "*serious and committed*" peers (who don't read a joke). To illustrate, we note that forty-five percent of our experiment subjects had not written any reviews in the past three months. They were likely to behave differently from the self-selected experienced evaluators on Yelp, who are themselves more likely to give an exaggerated rating in order to build a distinctive online identity (e.g., Dellarocas et al. 2010; Zeng and Wei 2013). Had these novice experiment subjects provided a rating as their experience and commitment evolved, they would likely have given a more corrective rating, which we believe (would skew) skews the data towards exaggerated ratings.

Yet another reason for the weaker correction may be due to the fact that the joke ratings have a rather larger variance, perhaps because of the very nature of joke appreciation. Thus, evaluators of jokes may not detect aberrations as well as evaluators of restaurants, and are less assertive when it comes to aberration correction. If this were true, we would expect a smaller aberration to result in an even weaker correction, and that is indeed what we find in the follow-up experiment we describe next.

5.2. Experiment 2 (A Follow-up Experiment)

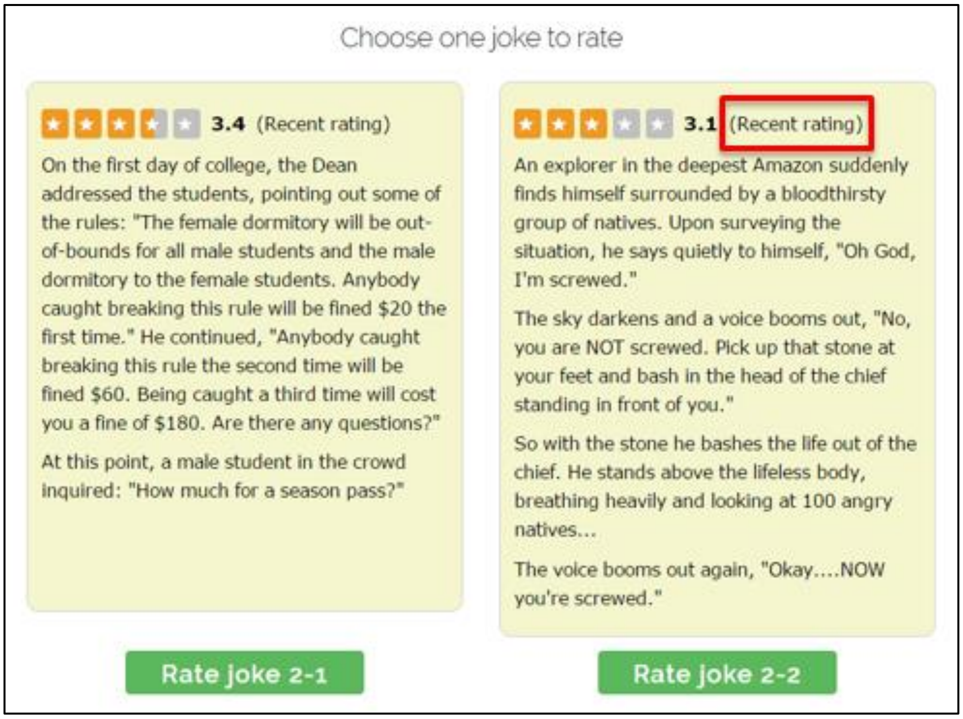
5.2.1. Experiment Design

The follow-up experiment aims to address two issues. First, we examine whether a smaller aberration is followed by a weaker correction. To do so, we set the size of the up and down treatments to 0.24 or 0.26 respectively.[11] Another issue that we hope to address in Experiment 2 is whether evaluators will behave differently when we frame the existing rating as a *recent* rating, which matches our observational study [12] (see Figure 5). The design of Experiment 2 is otherwise identical to that of Experiment 1.

---

[11] This also roughly matches the mean aberration sizes in the observational study.

[12] Experiment refers to existing ratings simply as "rating", which may be construed as an average rating.


Choose one joke to rate

3.4 (Recent rating)

On the first day of college, the Dean addressed the students, pointing out some of the rules: "The female dormitory will be out-of-bounds for all male students and the male dormitory to the female students. Anybody caught breaking this rule will be fined $20 the first time." He continued, "Anybody caught breaking this rule the second time will be fined $60. Being caught a third time will cost you a fine of $180. Are there any questions?"

At this point, a male student in the crowd inquired: "How much for a season pass?"

3.1 (Recent rating)

An explorer in the deepest Amazon suddenly finds himself surrounded by a bloodthirsty group of natives. Upon surveying the situation, he says quietly to himself, "Oh God, I'm screwed."

The sky darkens and a voice booms out, "No, you are NOT screwed. Pick up that stone at your feet and bash in the head of the chief standing in front of you."

So with the stone he bashes the life out of the chief. He stands above the lifeless body, breathing heavily and looking at 100 angry natives...

The voice booms out again, "Okay....NOW you're screwed."

Figure 2.5. Highlight Displayed Rating as Recent Rating

In Experiment 2, we recruited 401 subjects on Amazon Mechanical Turk. Using the same criteria as Experiment 1, we exclude 63 subjects and keep 338 for our analysis. Among usable samples, 54% are female. The summary statistics are quite similar to those in Experiment 1(see Table 8).

Table 2.8. Variable Definitions and Descriptive Statistics – Experiment 2

Variables	Mean	Std. Deviation	Min	Max	# Obs
Valence	0.909	1.092	-2.4	3	2,546
Up Treatment	0.203	0.402	0	1	2,546
DownTreatment	0.291	0.454	0	1	2,546
GenderFemale	0.555	0.497	0	1	2,546
AgeGroup	2.564	1.033	1	5	2,546
SkipSomeJokes	0.115	0.32	0	1	2,546
UpTreatment * SkipSomeJokes	0.026	0.159	0	1	2,546
DownTreatment * SkipSomeJokes	0.03	0.17	0	1	2,546

Variables Definitions are the same as those in Experiment 1.

5.2.2. Results

Table 9 reports the results from Experiment 2. We find that, across all three models, the coefficients of $UpTreatment$ are positive and significant, and the coefficients for $DownTreatment$ are negative and significant. Thus Experiment 2 replicates the qualitative results of Experiment 1. That said, compared with Experiment 1, the coefficients for the aberrations are smaller in sizes and less significant. We note that the interaction terms are no longer significant but the directions are the same as those in Experiment 1. Table 10 compares the correction percentages in the two experiments. As seen from this table, the correction percentages in Experiment 1 are larger than those in Experiment 2 across all models and treatment conditions. This suggests that the size of aberrations does matter, and people do respond to larger aberrations with stronger correction.

Table 2.9. Fixed-effect OLS Regression on Valence – Experiment 2

Independent Variables	**Model 1**	**Model 2**	**Model 3**
	Coefficient (SE)	**Coefficient (SE)**	**Coefficient (SE)**
UpTreatment	0.172* (0.068)	0.169* (0.068)	0.138+ (0.071)
DownTreatment	-0.159** (0.055)	-0.155** (0.055)	-0.143* (0.058)
GenderFemale		0.016 (0.043)	0.016 (0.043)
AgeGroup		0.002 (0.021)	0.001 (0.021)
SkipSomeJokes		0.142* (0.067)	0.118 (0.093)
UpTreatment * SkipSomeJokes			0.251 (0.170)
DownTreatment *SkipSomeJokes			-0.123 (0.160)
Constant	3.019*** (0.030)	2.988*** (0.068)	2.993*** (0.068)
Adjust R-squared	0.221	0.221	0.222
N	2,546	2,546	2,546

+ p<0.10, * p<0.05, ** p<0.01, *** p<0.001.

Table 2.10. Comparison of Average Correction Percentage across Models

		Average Correction Percentage	
		Experiment 1	**Experiment 2**
	Manipulation sizes:	+1, -1	+0.24, -0.26
Model 1	UpTreatment	63%	28%
	DownTreatment	74%	39%
Model 2	UpTreatment	63%	30%
	DownTreatment	74%	40%
Model 3	UpTreatment	64%	43%
	DownTreatment	76%	45%

6. Discussion and Implications

Motivated by the vital importance of online rating systems, we ask a distinct question of whether online rating systems are resilient to rating aberrations. We approach this question using a combination of observational data and experimental approaches. We first collect a large dataset on Yelp restaurant ratings and analyze the relationship between rating aberrations and subsequent rating valence. We then conduct two randomized online experiments on Amazon Mechanical Turk, where we administrate randomized aberrations of two different sizes. Overall, we find a tendency of new ratings to correct recent aberrations, and correction percentages ranging from about 63% for positive and 74% for negative aberration in the experiments to 102.8 % and 103.2% for positive and negative aberration respectively in the observational data. Our results show that the correction strength for positive aberrations is on par with those for negative ones. Furthermore, larger aberrations are followed by stronger corrections in both observational study and experimental studies. These results show that temporal aberrations in online ratings do not seem to cast a long shadow on future ratings, and the studied online rating systems are fairly robust to aberrations.

6.1. Academic and Practical Contributions

Our first contribution is to shine a light on an important yet under-addressed issue of reliability in online rating systems – the resistance to rating aberrations. We argue that aberration resistance is key to the long-term viability and trustworthiness of online rating systems. We contribute to this issue by developing the concept of rating aberrations, and examine their relationship with the valence of subsequent ratings. Overall, our study shows that online rating systems, at least the ones studied, seem to be fairly robust to rating aberrations. In the observational study of Yelp ratings, we observe a pattern of over-correction that could not be explained by "regression to the mean." To our knowledge, this is the first empirical evidence that online rating systems may be fairly resistant to rating aberrations.

On the issue of whether corrections for positive and negative aberrations are equal in strength, we reach a different conclusion than Muchnik et al. (2013), who study a similar issue in online voting. We show that the correction strength for positive aberrations is on par with that for negative ones. This is confirmed by both the observation study and the two experiments. The symmetry in corrective strength is good news for online rating systems and their users, because otherwise, the rating systems could become more biased in one direction. Our findings differ from Muchnik et al. (2013), who show that an artificial positive vote casts a long shadow but a negative does not. We argue that the difference in findings could be explained by the inherent greater significance of a negative vote than a positive vote. The different findings create an opportunity for future research to ascertain the causes of such a discrepancy.

Our findings hold important implications for business owners, consumers, and online rating platforms. In particular, our finding that rating aberrations do not appear to cast a long shadow on future ratings, is reassuring to consumers, who increasingly rely on such online systems. This may also help explain why we have relied on online ratings so much, despite the known representational biases of online ratings (Gao et al. 2012; Hu et al. 2009). Furthermore, the concepts and analytical techniques used in this study can be a basis for merchants, online rating platforms, and policy makers to develop monitors of rating aberrations and the resistance to aberrations at different online rating systems, so that warnings and corrective actions can be taken.

6.2. Limitations and Future Research

As the research on aberration resistance is still nascent, our study has a few limitations. We have found different corrective percentages in our online experiments and in our observational studies. We conjecture that this could be because evaluators are required to provide a rating in experiments whereas real-world evaluators are self-selected volunteers and thus could be more assertive and corrective. This discrepancy reflects an inherent challenge in designing experimental studies of rare behaviors: it may be difficult to obtain an adequate number of such behaviors if we replicate the rate at which they happen in the real world. We hope that future advances in experimental techniques for rare behaviors could provide further experimental tests of our observational findings.

In our observational study, we have relied on recent ratings for calculating rating aberrations, which is an approximation of what evaluators may see before they provide a new rating. As we have mentioned, Yelp, like many other rating systems, uses a reviewing sorting algorithm that considers not only recency, but also the popularity of reviews and the elite status of evaluators (Yelp Support Center), as well as giving users an option to sort reviews by recency or ratings alone. A more granular dataset such as one that documents click stream of evaluators could be useful for deriving a more accurate measurement of the observed rating aberrations.

Due to our data limitations, we have not been able to delineate the many possible mechanisms behind the observed corrective patterns. Nevertheless, we outline several such possible theoretical explanations and pave the way for future research on this important issue of aberration resistance.

For example, future studies could test what type of online users tend to be more corrective, how the corrective patterns differ among different stores, and how the design of online rating systems can shape the strength of aberration correction and therefore the long-term viability of rating systems. More laboratory or field experiments could also be designed to separate these potential mechanisms for aberration corrections. We believe that increased understanding of these issues can provide useful clues on how to design robust online rating systems.

Chapter 3. An Examination of Online Review Quality and Its Relationship with Helpfulness Votes

1. Introduction

User-generated online reviews of products and services have become a dominant source of information for consumers. According to a 2018 report by BrightLocal, about 86 percent of U.S consumers consult online reviews, among them, only 5% of consumers aged between 18 and 34 never read online reviews. On average, consumers read 10 online reviews before feeling able to make their decisions (BrightLocal 2018). Not all reviews are of equal quality. Some reviews are more informative and trustworthy than others. The extent to which consumers can benefit from consulting online reviews depends on the quality of online reviews. High-quality reviews help reduce information asymmetry between consumers and vendor (Pavlou et al. 2007) and thus help consumers make informative decisions. Low-quality reviews on the other hand can distract, burden, and mislead consumers.

Measuring review quality is difficult, however, because there is no agreed-upon concept of online review quality or an objective measurement. In practice, online review platforms and consumers often rely on helpfulness votes as an implicit substitute for review quality. Online platforms let consumer vote on whether a review is helpful. Helpfulness votes are typically displayed in one of the two ways: (1) the total number of helpfulness votes and (2) the percentage of helpfulness votes (if “not helpful” is also an option). These helpfulness votes are used by consumers and also factored into the review ranking and filtering algorithms by the online platforms.

Perhaps because of the prevalence of helpfulness votes in practice, the academic literature on online reviews has a predominantly focus on helpfulness votes and used them either as a convenient proxy for review quality or as an end in itself (Ghose and Ipeirotis 2011; Mudambi and Schuff 2010; Yin et al. 2014). There is a large literature devoting on deciphering helpfulness votes – i.e., what kinds of reviews tend to receive more helpfulness votes – and use these insights to guide review generation (Cao et al. 2011; Eslami et al. 2018; Mudambi and Schuff 2010; Salehan and Kim 2016; Yin et al. 2014).

The reliance on helpfulness votes as a proxy for review quality is not without controversy. Recent research suggests that such vote-based quality measurements may be biased, not fully- or even mis-communicate the true quality of reviews as voting can be affected by factors unrelated to review quality (Chen and Tseng 2011; Liu et al. 2007; Muchnik et al. 2013). For example, the more votes a review receives, the more authoritative it appears to readers, which in turn influences readers’ voting (Chen and Tseng 2011; Liu et al. 2007). Despite these concerns, there is not much research to systematically compare helpfulness votes and review quality or to remedy these issues.

A key step to address the above gaps in literature and practice is to clarify the concept of online review quality and propose a measurement of review quality. Only so, one can begin to document the relationship between review quality and helpfulness votes and study the antecedents and consequences of any discrepancy. In this paper, we begin to address this critical need by focusing on two issues: (1) develop the concept of online review quality and propose a measurement for it; (2) examine the relationship between helpfulness votes and review quality to gain insights on factors contributing to their discrepancies.

Developing a robust measurement of review quality is especially important for consumers and online review platforms for at least two reasons: first, it can be used to filter out lower-quality reviews that burden or mislead consumers; second, by beginning to reward high-quality reviews, the platform can encourage contribution of high quality reviews.

Perhaps the closest research is the prior work on information quality, developed in late 1990s in the context of organizations (DeLone and McLean 1992; Eppler and Wittig 2000; Lee et al. 2002; Wang and Strong 1996). Measurement of information quality may not transfer completely to online reviews because the latter is a specific kind of information with unique purpose of use. For example, accuracy, a key dimension in information quality, is not as relevant in the online review contexts because online reviews often involve a subjective evaluation that cannot be judged on the basis of right or wrong.

To address the first question, we adopt the Delphi method of ontology design. The Delphi method is a collaborative approach for ontology design where the goal is to reflect viewpoints and experiences of people who intentionally cooperate to produce the ontology (Gruber 1995). This approach has been widely used to develop concepts and measurements in Information Systems research (Chou et al. 2014; Holsapple and Joshi 2002; Okoli and Pawlowski 2004). We use this approach to develop a concept of online review quality and its measurement that reflect a collective view from diverse vantage points.

After developing the measurement of review quality, we used it to study a set of online reviews from Amazon with the goal of examining the relationship between helpfulness votes and review quality. We collect a set of online reviews from multiple product categories on Amazon. We hire Amazon Mechanical Turkers (AMTurkers) to rate the quality of reviews based on the measurement of review quality that we have developed. We then compare helpfulness votes with review quality.

Overall, we find that helpfulness votes are a poor indicator of review quality and statistically different from review quality. Specifically, shorter or negative reviews tend to receive more helpfulness votes relative to their review quality, and such an inflation tends to occur among low-quality reviews. In contrast, photo-augmented reviews (i.e. reviews with accompanying photos) tend to receive fewer helpfulness votes relative to their review quality. This deflation tends to occur among high-quality reviews.

2. Related Literature

2.1. Measurements of Information Quality

Literature has developed the measurements of information quality primarily in the context of organizations. It takes empirical approaches (Wang and Strong 1996; Zmud 1978) or based on existing literature (DeLone and McLean 1992; Goodhue 1995; Jarke and Vassiliou 1997; Lee et al. 2002) to develop the measurements. For example, Wang and Strong (1996) empirically identify dimensions from information consumers perspective. Several studies only focus on a few dimensions of information quality that can be objectively evaluated (Ballou and Pazer 1985; Wand and Wang 1996). For example, Wand and Wang (1996) consider four dimensions – correctness, unambiguous, completeness and meaningfulness, whose true values are available in the real world. Therefore, the corresponding values in a system can be identified by comparing with those in the real world.

These measurements of information quality have been used to assess review quality (Cheung et al. 2008; Park et al. 2007). As we discussed in the introduction, the measurements of information quality are primarily developed in the context of organizations (DeLone and McLean 1992; Eppler and Wittig 2000; Lee et al. 2002; Wang and Strong 1996), reviews as a specific kind of information have its unique purpose of use and quality dimensions. Therefore, we aim to develop a measurement for assessing review quality.

2.2. Helpfulness Votes of Online Review

The literature of online review predominantly focuses on helpfulness votes either as a convenience proxy for review quality or as a focus in itself (Forman and Ghose 2008; Ghose and Ipeirotis 2011; Mudambi and Schuff 2010; Yin et al. 2014). An abundance of research focuses on decoding the secret of helpfulness reviews – i.e., what kinds of reviews are helpful, by examining the association between features of a review's textual content and helpfulness votes (Cao et al. 2011; Ghose and Ipeirotis 2011; Mudambi and Schuff 2010; Yin et al. 2014). For example, subjectivity, readability, linguistic correctness, review extremity, review depth and length have a positive effect on helpfulness votes (Ghose and Ipeirotis 2011; Mudambi and Schuff 2010; Pan and Zhang 2011). Among these, several studies examine the effects of emotion and sentiment in reviews on helpfulness votes (Forman and Ghose 2008; Malik and Hussain 2017; Ullah et al. 2015; Yin et al. 2014). For example, Malik and Hussain (2017) shows that negative emotions (e.g., anxiety and sadness) have greater impact on helpfulness votes than positive emotions (e.g., trust, joy, and anticipation).

In this literature stream, a growing number of studies examine contextual factors, such as the product type and reviewer characteristics (Ghose and Ipeirotis 2011; Karimi et al. 2017; Lu et al. 2010; Mudambi and Schuff 2010; Pan and Zhang 2011). We add to this stream of research by examining the relationship between helpfulness votes and review quality.

3. Measurement Development

The collaborative approach to ontology design typically includes three development phases – i.e., preparation, anchoring, and iterative improvement (Holsapple and Joshi 2002), so does our development of the concept and measurement for online review quality.

Preparation. In the preparatory phase, we define design criteria (i.e., criteria used to evaluate the concept and measurement) and specify boundary conditions for the concept and measurement of online review quality. The design criteria we select for the definition of online review quality are completeness, correctness, conciseness, and clarity because they have been widely used for theory evaluation (Bacharach 1989; Kerlinger 1986). Two criteria are selected for evaluating the dimensions of online review quality: completeness (whether the dimension list is complete) and pertinency (i.e., whether a given dimension is pertinent to the concept of online review quality).

We apply the design criteria within three boundary conditions: *online boundary*, *measure boundary*, and *item boundary*. That is, our focus is on online (not offline) reviews, on measuring the quality of online review content (i.e., review title, rating, and content including photos and texts) rather than the quality of items being reviewed, and on goods (i.e., products and services, not people).

Anchoring. Anchoring is to seed a concept for collaborative design activity, and the seed serves an anchor to help focus the attention of collaborations. We draw on the concept of information quality to produce the initial (i.e., anchoring) concept of online review quality for seeding the collaborative effort.

Eppler and Wittig (2000) reviews the definitions of information quality and found that the most commonly used definitions of information quality are the following fours:

- "*Information quality can be defined as information that is fit for use by information consumers*" (Huang et al. 1999).
- "*Information quality is the characteristic of information to meet or exceed customer expectations*" (Kahn and Strong 1998).
- "*Quality information is information that meets specifications or requirements*" (Kahn and Strong 1998).
- "*Information quality is the characteristic of information to be of high value to its users*" (Lesca et al. 1995).

Users of online reviews are potential consumers who use the reviews to evaluate the item being reviewed and the concept of online review quality focuses on the quality of review content. Therefore, drawn on the four most commonly used definitions of information quality, we specify the anchoring definition of online review quality as "*the degree to which the content of an online review is valuable for potential consumers to evaluate the item.*"

Eppler and Wittig (2000) studies the series of dimensions in existing research of information quality, and states that only Wang and Strong (1996)'s dimensions provides both a theoretical foundation and practical applications. Based on Wang and Strong (1996)'s dimensions, we specify the anchoring dimensions as *relevant*, *trustworthy*, *objective*, *authoritative*, *original*, *comprehensible*, *organized*, *concise*, *informative*, *entertaining,* and *timely* (Table 1).

Table 3.1. Anchoring Dimensions of Online Review Quality

Dimension	**Description**
Relevant	The extent to which the information in a review is relevant to the item being evaluated.
Trustworthy	The extent to which the review is trustworthy.
Objective	The extent to which the review is unbiased (unprejudiced) and impartial.
Authoritative	The extent to which the review is authoritative.
Original	The extent to which the review is original.
Comprehensible	The extent to which the review is clear, without ambiguity, and easily comprehended.
Organized	The extent to which the review is well organized.
Concise	The extent to which the information in a review is compactly and concisely presented.
Informative	The extent to which the review is informative.
Entertaining	The extent to which the review is entertaining.
Timely	The extent to which the review is timely and up-to-date

We exclude *accessibility*, *security*, and *accuracy* in Wang and Strong (1996)'s dimensions of information quality from the anchoring dimensions of review quality because they are not applicable to online reviews – online reviews can be accessed by anyone who have a device connected to the Internet and are provided by voluntary consumers who do not have accurate information about the items (i.e., no accurate reviews), rather share their own subjective experience.

To tailor the dimensions in Wang and Strong (1996) for online reviews, we replace *reputation* (reputation of the information source and the information), *believability*, and *value-added* (give a competitive edge and value to information users) with authoritativeness, trustworthiness, and originality (*authoritative*, *trustworthy*, and *original* in Table 1) because authoritative and trustworthy reviews may be valuable for consumers to evaluate the item. We also combine *amount of information* and *completeness* (breadth, depth, and scope of information) as informativeness (*informative* in Table 1), and *interpretability* and *consistent representation* as comprehensibility (*comprehensible* in Table 1). We add entertainment (*entertaining* in Table 1) because consumers may also value hedonic metric while evaluating the quality of online reviews – entertaining reviews may be valuable to consumers (e.g., enjoy reading and easy to memorize the information in the reviews).

Iterative Improvement. In the phase of iterative improvement, a Delphi study is conducted to collect and integrate the views of collaborators (participants) about the concept, revise the concept to address collaborators' feedback, and iterate until consensus reached.

In the winter of 2017, 172 nonstudent researchers who had published in the conferences of *Association for Information Systems* (*the International Conference on Information Systems*, *the Americas Conference on Information Systems*, *Pacific Asia Conference on Information Systems*, *European Conference on Information Systems*) in the past five years are identified for participation on the panel. We conduct the Delphi study via a web-based application and invite the candidates via emails with links included to access the web-based application. Of the 172 candidates, 18 (10.5%) responded to the first round of the Delphi study. Among the 18 respondents, 1) one from Canada, two from mainland China, two from Hongkong of China, and the rest from the United States; 2) 17% have more than 10 years of research experience on online reviews, 33% have 2 or 3 years, and the rest have research experience on online reviews ranging from 4 to 8 years; and 3) 17% published more than 10 research articles (conference papers, journal papers, and book chapters) on online reviews; 50% published less than 3, and the rest published the number of research articles between 3 and 7.

In the first round of the Delphi study, participants are asked to rate the anchoring definition and dimensions of online review quality on the design criteria on a 5-point scale, along with elaborating their ratings and suggesting for improvement – i.e., rate the anchoring definition in terms of completeness, correctness, conciseness, and clarity (e.g., 1 – not at all clear, 5 – very clear), the anchoring dimension list in terms of completeness, and each dimension in terms of pertinency.

Based on the first-round ratings of the Delphi study (Appendix A reports the ratings), participants did not reach consensus because more than 30% participants rate 3 or lower on many design criteria (Green 1982). For the anchoring definition, participants suggest that the purpose of use for online reviews may go beyond *evaluate the item*, and consumers may also use online reviews to *gather information about the item*. In addition, participants also point out that online review quality is not only subjective (i.e., quality perceptions of online reviews can vary with consumer-specific tastes, purposes, and use cases), but also intersubjective because there can be agreement among the population of consumers in terms of their subjective evaluations. To convey this notion of "*commonality*", we use the term “*consumer population*” in the definition. We also dropped the word “*valuable*” from the definition as participants criticize it too vague. Based the above feedback, we revise the definition to “*the degree to which the content of an online review allows consumer population to gather information about and/or evaluate the item*.”

The feedback from participants indicates that there are overlaps and redundancies in our initial list of dimensions (e.g. between "*authoritative*", "*trustworthy*", and "*objective*"). To address this, we consolidate dimensions that are overlapping. In particular, we merge “*objective*,” "*authoritative*," and "*original*" into "*trustworthy*" with a revised description (Table 2). Similarly, we merge "*organized*" and "*concise*" into "*comprehensible*," and rename it "*well-written*."

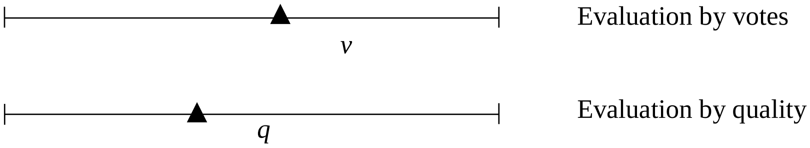
The feedback also indicates that a few dimensions are too broad (e.g., *informative*) and may be perceived too synonymous with the overall construct of "*quality*" (i.e., do not provide much differentiation from it). Therefore, we revise such dimensions to give them a narrower scope and to make sure that they are conceptually different from each other and from the overall "*quality*" construct. In particular, we revise "*informative*" to "*comprehensive*," and "*relevant*" to "*broadly applicable*." Table 2 shows the revised dimensions.

Table 3.2. Revised Dimensions of Online Review Quality

Dimension	**Description**
Relevant	The extent to which the review is broadly applicable to the consumer population.
Trustworthy	The extent to which the review is genuine, credible, and fair.
Comprehensive	The extent to which the review is detailed and thorough.
Well-written	The extent to which the review is well organized, clear, concise, and easy to understand.
Timely	The extent to which the review is timely and up-to-date.

In the second round of the Delphi study, 10 out of 18 respondents who participant in the first round of the Delphi study continue to participate in the second round. We provide participants feedback reflecting the combined views from the first round of the Delphi study. As the first round, the 10 participants are asked to rate the revised definition and dimensions of online review quality on the design criteria, elaborate their ratings, and suggest for improvement. Based on the second-round ratings on the design criteria (the ratings reported in Appendix B), 70 or higher percent of participants rate four or higher and the median at 4 or higher for all design criteria, thus the consensus is reached (Green 1982).

4. Theoretical Background and Hypotheses



We are interested the relationship between a review's helpfulness votes and its quality. To capture the discrepancy between the two, we study the difference between the evaluation of a review by helpfulness votes (i.e., bias of helpfulness votes), v, and that by quality, q, provided that both are normalized to the same scale. We say a review has an inflated evaluation by helpfulness votes if its normalized evaluation by votes is higher than that by quality (i.e. $v > q$). Conversely, we say the review has a deflated evaluation by helpfulness votes.

Theory of selective attention suggests that human beings have only a limited capacity to process information, therefore, we selectively attend to certain stimuli and inputs while ignoring others, in order to prevent the information-processing system from becoming overloaded (Broadbent 1958; Driver 2001; Treisman 1969). Attention is also a limited resource, selective attention allows us to allocate our available reserves among inputs and stimuli (Driver 2001; Kanfer and Ackerman 1989). Literature has showed that what inputs are selected is determined by our arousal level, which directs our awareness and attention to inputs associated with arousal (Deutsch and Gerard 1955; Yechiam and Hochman 2013).

Voters of online reviews are causal evaluators – they may selectively attend to certain reviews that arouse them. Therefore, their judgement may be more emotionally driven, less deliberate and thoughtful.

Length Bias

Because voters of online review are casual, ad hoc evaluators, they are very sensitive to evaluative cost. Length of reviews adds to such cost – longer reviews require more time to read the reviews and more cognitive effort to understand the information in the reviews. An abundant amount of evidence supports that humans selects the inputs based on the physical attributes (e.g., color, length, size) (Broadbent 1958; Driver 2001; Treisman 1969). Research has shown that shorter questionnaires are associated with high response rate as they require less response cost (Hedlin et al. 2005; Salisbury et al. 2005). Similarly, shorter reviews require less time, cognitive effort, and evaluation cost, therefore, they are more likely to be selected by voters.

More voters selectively attend to shorter reviews, leading shorter reviews to receive more votes. While both up and down voting are possible for shorter reviews, most platforms only offer upvoting. With this, shorter reviews, due to being able to attract more attentions, tend to receive more upvotes. Thus, we hypothesize that:

Hypothesis 1 (H1): *A shorter review tends to receive a more inflated evaluation by helpfulness votes (relative to its quality).*

Negativity Bias

Low-rating (negative) reviews imply losses – i.e., authors of negative reviews have negative experience (loss) with the items being reviewed. Loss increases attention allocated to the inputs associated with the losses (Taylor 1991). Even if people who do not weight loss more than gain, they still pay more attention to loss because loss leads to a momentary increase in arousal, which directs attention to the stimuli and inputs associated with the losses (Pribram and McGuinness 1975; Yechiam and Hochman 2013). In other words, increased arousal following losses leads to more attention even without loss aversion. Experiencing losses or potential losses facilitate an increase in the attention (Yechiam and Hochman 2013).

Voters of online reviews are potential consumers, who read and consult online reviews with an aim to gather information about and/or evaluate the item. They tend to selectively attend to negative reviews because negative reviews are associated with losses and the losses increase arousal in voters, who then direct their attention to negative reviews in response to the arousal, even if the voters do not intend to purchase the item being reviewed – i.e., there is no loss associated with the voters. In addition, voters with intention to purchase, tend to pay more attention to negative reviews because they have potential of experiencing losses – i.e., authors of negative reviews have negative experience (loss) with the items, and they have the potential of experiencing the same loss if they purchase the items.

Recent research has provided empirical evidence that users tend to pay more attention to negative reviews because negative reviews appear more professional, trustworthy, valuable, and informative even if they are actually not (Anderson and Simester 2014; Linton et al. 2017). Therefore, we expect that

Hypothesis 2 (H2): *A low-rating review tends to receive a more inflated evaluation by helpfulness votes (relative to its quality)*.

Photo-augmented Bias

Literature suggests that visual information (e.g., imagery, video, and photo) tends to stimulate more affections and emotions than verbal information (e.g., language, and auditory and written narrative and words) (Clark and Paivio 1991; Sadoski et al. 2000). The research has shown that compared to equivalent verbal information, visual information elicit higher emotional engagement – i.e., people are more engaging and interesting in visual information (Koehler et al. 2005) and tend to collect more information via visual channel when both visual and verbal information are present (Birdwhistell 2010).

Photo-augmented reviews are visual information, and tend to attract more voters of online reviews because photo-augmented reviews lead to more emotional arousal in voters, which in turns directs more voters' attention to photo-augmented reviews. Hence, we offer the following hypothesis:

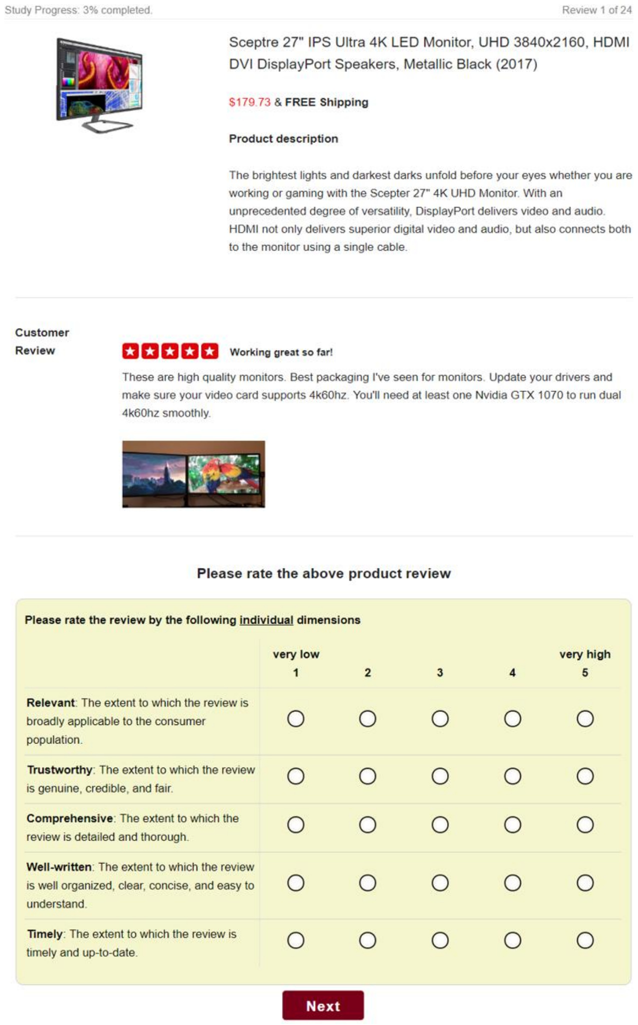
Hypothesis 3 (H3): *A photo-augmented review tends to receive a more inflated evaluation by helpfulness votes (relative to its quality)*.

5. Methodology

To examine the relationship between review quality and helpfulness votes, we first gather a list of 3,256 products from four categories (i.e., mattresses, sneakers, monitors, and books) on Amazon on October 28, 2018. The aforementioned categories cover both experience goods (the first two) and search goods, and high- (e.g., monitor) and low-priced goods (e.g., books). Among the 3,256 products, we randomly sample 6 products for each category, resulting in 24 products in total. For each of the 24 products, we collect product information (e.g., pictures, price, description, average rating, sellers) and all of reviews (e.g., review title, helpful votes, date post, rating, review content, reviewer).

For each of the 24 products, we classify the reviews according to the length (long and short), the number of helpfulness votes (three levels), and age (two levels). By this classification scheme, each review belongs to one of the 12 groups based on its length, age, and helpfulness votes. We randomly sample 1 review from each of the 12 groups. The stratified random sampling is adopted to ensure the subsample has a balanced composition. We obtain a total of 274 reviews (noting that some groups are empty).

To obtain the quality of the 274 reviews, we hire Amazon Mechanical Turkers (AMTurkers) to rate the review quality using the measurement of online review quality we develop. While rating review quality, AMTurkers are provided with the definitions of dimensions, along with product information and review content including review title, rating, texts, and photos (see Figure 1). Information of reviewers is not provided as it may confound the evaluation of review quality. We randomly assign each AMTurker 24 reviews, one from each product (with few exceptions where there are empty groups). The order of reviews (products) is also randomized. We choose the number of tasks such that each review is evaluated three times. To ensure quality of responses, we select only U.S. Turkers with over 95% task acceptance rates and at least 50 completed assignments. They are paid $3 for participating in our study.



Study Progress: 3% completed.

Review 1 of 24

Sceptre 27" IPS Ultra 4K LED Monitor, UHD 3840x2160, HDMI DVI DisplayPort Speakers, Metallic Black (2017)

$179.73 & **FREE Shipping**

Product description

The brightest lights and darkest darks unfold before your eyes whether you are working or gaming with the Scepter 27" 4K UHD Monitor. With an unprecedented degree of versatility, DisplayPort delivers video and audio. HDMI not only delivers superior digital video and audio, but also connects both to the monitor using a single cable.

Customer Review

★★★★★ **Working great so far!**

These are high quality monitors. Best packaging I've seen for monitors. Update your drivers and make sure your video card supports 4k60hz. You'll need at least one Nvidia GTX 1070 to run dual 4k60hz smoothly.

Please rate the above product review

Please rate the review by the following individual dimensions

	very low 1	2	3	4	very high 5
Relevant: The extent to which the review is broadly applicable to the consumer population.	○	○	○	○	○
Trustworthy: The extent to which the review is genuine, credible, and fair.	○	○	○	○	○
Comprehensive: The extent to which the review is detailed and thorough.	○	○	○	○	○
Well-written: The extent to which the review is well organized, clear, concise, and easy to understand.	○	○	○	○	○
Timely: The extent to which the review is timely and up-to-date.	○	○	○	○	○

Figure 3.1. An example of Rating a Review's Quality

6. Results

We first conduct principal-component factor analysis to see whether all dimensions load into the same factor. Table 3 reports the results. The results show that all dimensions load into the same factor with high factor loadings (relevant = 0.915, trustworthy = 0.914, comprehensive = 0.899, well-written = 0.919, timely = 0.905, respectively), suggesting the five dimensions all are highly correlated with one latent factor, i.e., review quality. To obtain the quality of each review, we first average three ratings of each dimension, and then calculate the sum of weighed dimensions.

Table 3.3. Factor loading

Dimensions	**Factor1 (weight)**
Relevant	0.915
Trustworthy	0.914
Comprehensive	0.899
Well-written	0.919
Timely	0.905

Because we are interested in the relationship between review quality and helpfulness votes, we create our dependable variable, *bias*, as the difference between the normalized helpfulness votes and the normalized review quality. This variable captures the extent to which helpfulness votes deviates from review quality.

Before conducting regression analyses, we plot the relationship between the normalized logarithm helpfulness votes and the normalized review quality for the 274 reviews. Figure 2 shows that not all reviews fall in Quadrant I and III and some reviews fall in Quadrant II (reviews with inflated evaluations by helpfulness votes) and IV (reviews with deflated evaluations by helpfulness votes), suggesting that the difference (*bias*) exists.

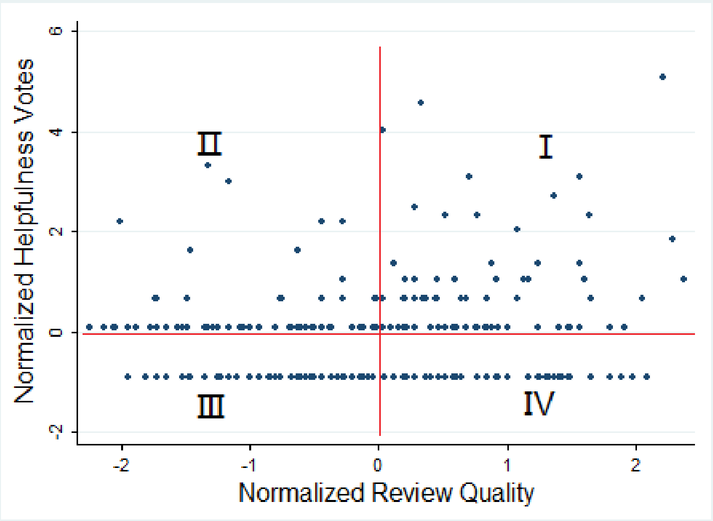


Figure 2.2. Relationship between Helpfulness Votes and Review Quality

We conduct regression analyses to further test our hypotheses and our unit of analysis is at the review level. Table 4 provides definitions of our variables, including review age (*LogAge*), length (*LogLength*), rating (*Rating*), the rank of the review author (*LogRank*), whether the review has accompanying with photos (*WithPhoto*) and whether the review author purchased the item being reviewed (*VerifiedPurchase*).

Table 3.4. Variable Definitions and Descriptive Statistics

Variables	Definition	Mean	Std. Dev	Min	Max	N
Bias_i	The difference between normalized log # helpfulness votes and normalized quality of review i	0	1.34	-2.98	4.65	274
NormQuality_i	Normalized quality of review i	0	1.00	-2.25	2.37	274
NormVotes_i	Normalized log # helpfulness votes of review i	0	1.00	-0.89	5.09	274
LogLength_i	Log length (in bytes) of review i	4.76	1.28	2	8	274
Rating_i	Rating of review i	4.23	1.25	1	5	274
WithPhoto_i	Whether review i is with photos: yes 1; otherwise 0	0.06	0.24	0	1	274
LogAge_i	Log days between review i posted and collected (October 28, 2018)	5.46	1.41	2	9	274
$\text{VerifiedPurchase}_i$	Whether review i's author bought the item being reviewed: yes 1; otherwise 0	0.90	0.30	0	1	274
LogRank_i	Log rank[1] of review i's author	1.52	1.80	7	18	274

[1] A reviewer's rank is determined by the overall helpfulness of all his reviews, factoring in the number of reviews he have written.

We estimate an OLS model and the results are reported in Model 1 of Table 5. Our results show the coefficients of *LogLength* and *Rating* are negative and significant (*Coeff* = -0.365, p < 0.001; *Coeff* = -0.202, *p* = 0.004; respectively), suggesting that shorter and low-rating (negative) reviews tend to receive more helpfulness votes relative to their quality (i.e., helpfulness votes bias against long and positive reviews). Therefore, Hypothesis 1 and 2 are supported.

Surprisingly, the coefficient of *WithPhoto* is negative and significant (*Coeff* = -1.988, $p < 0.001$), demonstrating that a photo-augmented review tends to receive fewer helpfulness votes relative to its quality (i.e., helpfulness votes bias against photo-augmented reviews). Therefore, Hypothesis 3 is not supported. One possible explanation is that consumers tend to evaluate such reviews primarily based on photos because visual information is more engaging and elicits more emotional responses (Birdwhistell 2010; Koehler et al. 2005; Sadoski et al. 2000). This, combined with the fact that photos are unable to communicate certain information about a product (e.g., service, sound quality), leads to fewer helpfulness votes.

Table 3.5. OLS Regression Results

Independent Variables	**Model 1**	**Model 2**
	Coefficient (Robust SE)	**Coefficient (Robust SE)**
$LogLength_i$	-0.365***	-0.366***
	(0.075)	(0.074)
$Rating_i$	-0.202**	-0.205**
	(0.069)	(0.072)
$WithPhoto_i$	-1.988***	-1.983***
	(0.191)	(0.191)
$LogAge_i$	-0.093+	-0.092+
	(0.049)	(0.049)
$VerifiedPurchase_i$	-0.020	-0.021
	(0.304)	(0.303)
$LogRank_i$	-0.026	-0.026
	(0.049)	(0.049)
$PercentNegativeRatings_i$		-0.001
		(0.005)
Constant	3.637***	3.667***
	(1.036)	(1.038)
Log-likelihood	-418.45	-418.43
Adjusted R-squared	0.288	0.285
N	274	274

Notes: ① DV = the difference between the normalized log helpfulness votes and the normalized review quality of review i (*$bias_i$*). ② The value in parenthesis is robust standard errors. + p < 0.10, *p < 0.05, **p < 0.01, ***p < 0.001.

One concern could be that the negativity bias (i.e., helpfulness votes favor negative reviews) is because of rarity negative reviews, which may suggest that negative reviews are more diagnostic and deserve to be viewed as higher quality. To address this concern, we control for *PercentNegativeRatings*, defined as the percent of negative reviews (reviews' rating with 1 standard deviation from the mean) among all reviews of the product. The OLS results are reported in Model 1 of Table5. The results show that after controlling for *PercentNegativeRatings*, the coefficient for Rating is still significant, suggesting that the negativity bias may not be a result of the scarcity of negative reviews.

To further understand the relationship between helpfulness votes and review quality, we split our data set (i.e., the 274 reviews) into two based on normalized review quality (*NormQuality*) – one data set consists of low-quality reviews (*NormQuality* < 0), whereas the other, high-quality reviews (*NormQuality* > 0). We have used different cutoffs (0, 0.1, -0.1, 0.2, -0.2) of review quality and the results are similar. As seen in Table 6, the coefficients of *LogLength* and *Rating* are negative and significant (*Coeff* = -0.358, *p* < 0.001; *Coeff* = -0.334, *p* <0.001; respectively) in Model 1, suggesting that length- and rating-based biases tends to occur among low-quality reviews. The coefficient of *WithPhoto* is negative and significant (*Coeff* = -1.800, *p* < 0.001) in Model 2, suggesting that high-quality reviews tend to suffer from the photo-based bias.

Table 3.6. OLS Regression Results on Low-quality and High-quality Reviews

Independent Variables	**Model 1 (Low-quality Reviews)**	**Model 2 (High-quality Reviews)**
	Coefficient (Robust SE)	**Coefficient (Robust SE)**
$LogLength_i$	-0.358***	0.177
	(0.077)	(0.143)
$Rating_i$	-0.334***	-0.045

	(0.081)	(0.082)
WithPhoto$_i$	-[1]	-1.800***
	-	(0.228)
LogAge$_i$	-0.007	-0.144*
	(0.054)	(0.072)
VerifiedPurchase$_i$	-0.700	0.254
	(0.565)	(0.324)
LogRank$_i$	-0.056	0.018
	(0.062)	(0.061)
PercentNegativeRatings$_i$	0.001	0.005
	(0.007)	(0.006)
Constant	5.095***	-1.048
	(1.244)	(1.491)
Log-likelihood	-175.53	-208.68
Adjusted R-squared	0.248	0.162
N	139	135

Notes: ① DV = the difference between the normalized log helpfulness votes and the normalized review quality of review i ($bias_i$). ② The value in parenthesis is robust standard errors. + $p < 0.10$, *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

[1] *WithPhoto* is not estimated because there are no reviews with photo in low-quality reviews.

7. Discussion and Implications

We develop a theoretically-grounded concept and measurement for review quality, and then investigate the relationship between helpfulness votes (a commonly used quality indicator) and review quality. We find that helpfulness votes are not a good indicator of review quality because helpfulness votes are biased and statistically different from review quality. Review length, rating, and photos all contribute to the difference and bias – shorter or negative (low-rating) reviews induce inflation evaluations (i.e., length and negativity biases), whereas photo-augmented reviews lead to deflation evaluations by helpfulness votes relative to their review quality (i.e., photo-augmented bias). In addition, we find that length and negativity biases tend to occur among low-quality reviews, whereas photo-augmented bias tends to occur among high-quality reviews.

7.1. Contributions to the Literature

Our research contributes to the literature in three ways. Our research makes a unique contribution to the information literature by developing a concept and measurement for the quality of online reviews, a unique type of information. Our concept and measurement provide theoretical ground for studies that involve review quality assessment. In addition, we find that the concept and measurement of review quality is very different from those of information quality, which provides support for a need of developing the concept and measurement for review quality.

Second, we contribute to the literature on the measurement of review quality. We find that helpfulness votes, a commonly used quality measure in prior research, is biased and misrepresents review quality. Based on our findings, helpfulness votes should not be used to approximate review quality.

Finally, we contribute to the literature on biases in judgement and evaluation by identifying and quantifying biases of helpfulness votes in the assessment of review quality. We find three biases in helpfulness votes as a measure for review quality (i.e., length, negativity, and photo-augmented biases). Moreover, low-quality reviews suffer from length and negativity biases, while high-quality reviews suffer from photo-augmented bias. These findings provide insights on studying biases in judgement and evaluation in other contexts.

7.2. Managerial Implications

Our first managerial implication is that online review platforms can use our measurement to filter out low-quality reviews and sift out high-quality ones, and present those high-quality reviews to assist consumers for their decision making. Our findings suggest that a high-quality review has to be relevant, trustworthy, comprehensive, well-written, and timely. Online review platforms can highlight and recommend reviews that satisfy the five quality dimensions (i.e., high-quality reviews) to their users, which will help online review platforms maintain current users and attract new users because it’s easy for users of the platform to find quality reviews for their decision making.

Online review platforms can also use our measurement to guide and advise review generation. For example, if a reviewer tends to write irrelevant content for his reviews, the platform may advise him to write more relevant content and how. With more quality reviews generated, online review platforms can increase their use base.

Our findings suggest that helpfulness votes are biased and statistically different from review quality. Online review platforms offering helpfulness votes may need rethink this design because helpfulness votes mislead users – shorter or negative reviews tend to receive more helpfulness votes relative to their quality and these length and negativity biases tend to occur among low-quality reviews, whereas photo-augmented reviews tend to receive fewer helpfulness votes relative to their quality and this photo-augmented bias tends to occur among high-quality reviews. Consequently, online review platforms likely loss users due to the misleading helpfulness votes – users leave the platform because of their inability to find quality reviews for their decision making; reviewers whose reviews receive fewer helpfulness votes relative to their quality (e.g., photo-augmented reviews) leave the platform because their reviews are unfairly voted.

7.3. Limitations and Future Research

A few limitations of this study are worth nothing, which provide directions for future research. First, while we use stratified random sampling to obtain our date set, our findings should be further tested using a larger data set. Second, we investigate whether helpfulness votes are a biased indicator for review quality. Future research can be extended to study other quality indicators such as expert reviews. Third, we identify three biases on Amazon reviews. It is worthwhile to identify other biases in other settings. For example, for online review platforms with social networking (e.g., Yelp), reviewers’ social networks (e.g., friends) may contribute to bias in helpfulness votes.

7.4. Concluding Remarks

The issue of review quality is an important one because it is useful for finding and creating high-quality reviews that better assist consumers in their decision making. This study develops a measurement of review quality and examine whether helpfulness votes is a good indicator of review quality. Our results indicate reviews that helpfulness votes are not good indicators of review quality and we discover several biases in helpfulness votes that are worthy for further attention.

Bibliography

Anderson, C., and Kilduff, G. J. 2009. "The Pursuit of Status in Social Groups," *Current Directions in Psychological Science* (18:5), pp. 295–298.

Anderson, C., and Shirako, A. 2008. "Are Individuals' Reputations Related to Their History of Behavior?," *Journal of Personality and Social Psychology* (94:2), p. 320.

Anderson, E., and Simester, D. 2014. "Reviews without a Purchase: Low Ratings, Loyal Customers, and Deception," *Journal of Marketing Research* (51:3), pp. 249–269.

Andreoni, J., and Miller, J. 1996. *Giving According to GARP: An Experimental Study of Rationality and Altruism*, Social Systems Research Institute, University of Wisconsin.

Bakshy, E., Rosenn, I., Marlow, C., and Adamic, L. 2012. "The Role of Social Networks in Information Diffusion," in *WWW '12 Proceedings of the 21st International Conference on World Wide Web*, Lyon, France: ACM, pp. 519–528.

Bapna, R., and Umyarov, A. 2015. "Do Your Online Friends Make You Pay? A Randomized Field Experiment on Peer Influence in Online Social Networks," *Management Science* (61:8), pp. 1902–1920.

Bond, R., Fariss, C., Jones, J., Kramer, A., and Marlow, C. 2012. "A 61-Million-Person Experiment in Social Influence and Political Mobilization," *Nature* (489:7415), pp. 295–298.

BrightLocal. 2016. "Local Consumer Review Survey." (https://www.brightlocal.com/learn/local-consumer-review-survey).

Burtch, G., Hong, Y., Bapna, R., and Griskevicius, V. 2018. "Stimulating Online Reviews by Combining Financial Incentives and Social Norms," *Management Science* (64:5), pp. 2065–2082.

Cao, Q., Duan, W., and Gan, Q. 2011. "Exploring Determinants of Voting for the 'Helpfulness' of Online User Reviews: A Text Mining Approach," *Decision Support Systems* (50:2), pp. 511–521.

Chen, Y., Harper, M., Konstan, J., and Li, S. X. 2010. "Social Comparisons and Contributions to Online Communities: A Field Experiment on MovieLens," *American Economic Review* (100:4), American Economic Review, American Economic Association, pp. 1358–1398.

Constant, D., Sproull, L., and Kiesler, S. 1996. "The Kindness of Strangers: The Usefulness of Electronic Weak Ties for Technical Advice," *Organization Science* (7:2), pp. 119–135.

Crowston, K., Wei, K., Li, Q., and Howison, J. 2006. "Core and Periphery in Free/Libre and Open Source Software Team Communications," in *System Sciences, 2006. HICSS'06. Proceedings* *of the 39th Annual Hawaii International Conference,* IEEE, p. 118a.

Dahlander, L., and Frederiksen, L. 2012. "The Core and Cosmopolitans: A Relational View of Innovation in User Communities," *Organization Science* (23:4), pp. 988–1007.

Danescu-Niculescu-Mizil, C., West, R., Jurafsky, D., and Potts, C. 2013. "No Country for Old Members : User Lifecycle and Linguistic Change in Online Communities," in *Proceedings of the 22nd International Conference on World Wide Web,* pp. 307–318.

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), pp. 127–158.

Dellarocas, C., and Narayan, R. 2006. "A Statistical Measure of a Population's Propensity to Engage in Post-Purchase Online Word-of-Mouth," *Statistical Science* (21:2), pp. 277–285.

Dellarocas, C., Zhang, X., and Awad, N. 2007. "Exploring the Value of Online Product Reviews in Forecasting Sales: The Case of Motion Pictures," *Journal of Interactive Marketing* (21:4), pp. 23–45.

Dennis, A. R., Pootheri, S. K., and Natarajan, V. L. 1998. "Lessons from the Early Adopters of Web Groupware," *Journal of Management Information Systems* (14:4), pp. 65–86.

Dewan, S., Ho, Y., and Ramaprasad, J. 2017. "Popularity or Proximity : Characterizing the Nature of Social Influence in an Online Music Community," *Information Systems Research* (28:1), pp. 117–136.

Dey, E. L. 1997. "Undergraduate Political Attitudes," *The Journal of Higher Education* (68:4), pp. 398–413.

Dichter, E. 1966. "How Word-of-Mouth Advertising Works," *Harvard Business Review* (44:6), pp. 147–160.

Donath, J. 2002. "Identity and Deception in the Virtual Community," *Communities in Cyberspace*, Routledge.

Duan, W., Gu, B., and Whinston, A. 2008. "Do Online Reviews Matter?—An Empirical Investigation of Panel Data," *Decision Support Systems* (45:4), pp. 1007–1016.

Forman, C., and Ghose, A. 2008. "Examining the Relationship between Reviews and Sales: The Role of Reviewer Identity Disclosure in Electronic Markets," *Information Systems Research* (19:3), pp. 291–313.

Fortune. 2016. "A Lack of Online Reviews Could Kill Your Business," *Fortune*. (http://fortune.com/2016/10/23/online-reviews-business-marketing).

Ghasemkhani, H., Kannan, K., and Khern-am-nuai, W. 2016. "Extrinsic versus Intrinsic Rewards to Participate in a Crowd Context: An Analysis of a Review Platform," *Working Paper*.

Ghose, A., and Ipeirotis, P. 2011. "Estimating the Helpfulness and Economic Impact of Product Reviews: Mining Text and Reviewer Characteristics," *IEEE Transactions on Knowledge and Data Engineering* (23:10), pp. 1498–1512.

Goes, P. B., Guo, C., and Lin, M. 2016. "Do Incentive Hierarchies Induce User Effort? Evidence from an Online Knowledge Exchange," *Information Systems Research* (27:3), pp. 497–516.

Goes, P. O., Lin, M., and Yeung, C. A. 2014. "'Popularity Effect' in User-Generated Contents: Evidence from Online Product Reviews," *Information Systems Research* (25:2), pp. 222–238.

Harbaugh, W., Mayr, U., and Burghart, D. 2007. "Neural Responses to Taxation and Voluntary Giving Reveal Motives for Charitable Donations," *Science* (316:5831), pp. 1622–1625.

Hardy, and Van Vugt, M. 2006. "Nice Guys Finish First: The Competitive Altruism Hypothesis," *Personality and Social Psychology Bulletin* (32:10), pp. 1402–1413.

Hiltz, S., and Wellman, B. 1997. "Asynchronous Learning Networks as a Virtual Classroom," *Communications of the ACM* (40:9), pp. 44–49.

Huang, N., Hong, Y., and Burtch, G. 2016. “Social Network Integration and User Content Generation: Evidence from Natural Experiments,” *MIS Quarterly* (Forthcoming).

Isaac, R., Walker, J., and Williams, A. 1994. “Group Size and the Voluntary Provision of Public Goods: Experimental Evidence Utilizing Large Groups,” *Journal of Public Economics* (54:1), pp. 1–36.

Jabr, W., Mookerjee, R., Tan, Y., and Mookerjee, V. 2014. “Leveraging Philanthropic Behavior for Customer Support: The Case of User Support Forums,” *MIS Quarterly* (38:1), pp. 187–208.

Jones, C., Hesterly, W., and Borgatti, S. 1997. “A General Theory of Network Governance: Exchange Conditions and Social Mechanisms,” *Academy of Management Review* (22:4), pp. 911–945.

Kaplan, M., and Miller, C. 1987. “Group Decision Making and Normative versus Informational Influence: Effects of Type of Issue and Assigned Decision Rule.,” *Journal of Personality and Social Psychology* (53:2), p. 306.

King, G., and Zeng, L. 2001. “Logistic Regression in Rare Rvents Data,” *Political Analysis* (9), pp. 137–163.

King, G., and Zeng, L. 2002. “Estimating Risk and Rate Levels, Ratios, and Differences in Case-Control Studies,” *Statistics in Medicine* (21), pp. 1409–1427.

Kumar, N., and Benbasat, I. 2006. “The Influence of Recommendations and Consumer Reviews on Evaluations of Websites,” *Information Systems Research* (17:4), pp. 425–439.

Lampel, J., and Bhalla, A. 2007. “The Role of Status Seeking in Online Communities: Giving the Gift of Experience,” *Journal of Computer-Mediated Communication* (12:2), pp. 434–455.

Lee, G. M., Qiu, L., and Whinston, A. B. 2016. “A Friend Like Me: Modeling Network Formation in a Location-Based Social Network,” *Journal of Management Information Systems* (33:4), pp. 1008–1033.

Lee, Tan, Y., and Hosanagar, K. 2015. “Do Ifollow My Friends or the Crowd? Information Cascades in Online Movie Rating,” *Management Science* (61:9), pp. 2241–2258.

Levina, N., and Arriaga, M. 2014. “Distinction and Status Production on User-Generated Content Platforms: Using Bourdieu’s Theory of Cultural Production to Understand Social Dynamics in Online Fields,” *Information Systems Research* (25:3), pp. 468–488.

Liu, D., Brass, D., Lu, Y., and Chen, D. 2015. “Friendships in Online Peer-to-Peer Lending: Pipes, Prisms, and Relational Herding,” *MIS Quarterly* (39:3), pp. 729–742.

Lu, Y., Tsaparas, P., Ntoulas, A., and Polanyi, L. 2010. “Exploiting Social Context for Review Quality Prediction,” in *Proceedings of the 19th International Conference on World Wide Web*, ACM, pp. 691–700.

Ludford, P., Cosley, D., and Frankowski, D. 2004. “Think Different: Increasing Online Community Participation Using Uniqueness and Group Dissimilarity,” in *CHI ’04 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Vienna, Austria: ACM, pp. 631–638.

Ma, X., Khansa, L., Deng, Y., and Kim, S. S. 2013. “Impact of Prior Reviews on the Subsequent Review Process in Reputation Systems,” *Journal of Management Information Systems* (30:3), pp. 279–310.

McGrath, J. 1984. *Groups: Interaction and Performance,* Englewood Cliffs, NJ: Prentice-Hall.

Mesch, G., and Talmud, I. 2006. “The Quality of Online and Offline Relationships: The Role of Multiplexity and Duration of Social Relationships,” *The Information Society* (22:3), pp. 137–148.

Moe, and Schweidel. 2012. “Online Product Opinions: Incidence, Evaluation, and Evolution,” *Marketing Science* (31:3), pp. 372–386.

Moretti, E. 2011. "Social Learning and Peer Effects in Consumption: Evidence from Movie Sales," *The Review of Economic Studies* (78:1), pp. 356–393.

Mudambi, S., and Schuff, D. 2010. "What Makes a Helpful Review? A Study of Customer Reviews on Amazon. Com," *MIS Quarterly*, pp. 185–200.

Mutter, T., and Kundisch, D. 2014. "Behavioral Mechanisms Prompted by Badges: The Goal-Gradient Hypothesis," in *Proceedings of the International Conference on Information Systems*, Auckland, New Zealand.

Nam, K. K., Ackerman, M. S., and Adamic, L. A. 2009. "Questions in, Knowledge in? A Study of Naver's Question Answering Community," in *CHI '09 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Boston, MA, USA: ACM, pp. 779–788.

Otterbacher, J. 2009. "'Helpfulness' in Online Communities: A Measure of Message Quality," *SIGCHI Conf. Human Factor Comput. Systems*, pp. 955–964.

Pan, Y., and Zhang, J. Q. 2011. "Born Unequal: A Study of the Helpfulness of User-Generated Product Reviews," *Journal of Retailing* (87:4), pp. 598–612.

Paolillo, J. 2008. "Structure and Network in the YouTube Core," in *Hawaii International Conference on System Sciences, Proceedings of the 41st Annual*, IEEE.

Pavlou, P., Liang, H., and Xue, Y. 2007. "Understanding and Mitigating Uncertainty in Online Environments: A Principal-Agent Perspective," *MIS Quarterly* (31:1), pp. 105–136.

Qiu, L., Shi, Z., and Whinston, A. 2017. "Learning from Your Friends ' Check-Ins : An Empirical Study of Location-Based Social Networks," *Information Systems Research* (Forthcoming).

Ren, Y., Kraut, R., and Kiesler, S. 2007. "Applying Common Identity and Bond Theory to Design of Online Communities," *Organization Studies* (28:3), pp. 377–408.

ReviewMeta.com. 2016. "Analysis of 7 Million Amazon Reviews: Customers Who Receive Free or Discounted Item Much More Likely to Write Positive Review." (https://reviewmeta.com/blog/analysis-of-7-million-amazon-reviews-customers-who-receive-free-or-discounted-item-much-more-likely-to-write-positive-review).

Ridings, C., and Gefen, D. 2004. "Virtual Community Attraction: Why People Hang out Online," *Journal of Computer-Mediated Communication* (10:1), p. JCMC10110.

Rui, H., and Whinston, A. 2012. "Information or Attention? An Empirical Study of User Contribution on Twitter," *Information Systems and E-Business Management* (10:3), pp. 309–324.

Sinha, R., and Swearingen, K. 2001. "Comparing Recommendations Made by Online Systems and Friends.," *DELOS Workshop: Personalisation and Recommender Systems in Digital Libraries* (106).

Sridhar, S., and Srinivasan, R. 2012. "Social Influence Effects in Online Product Ratings," *Journal of Marketing* (76:5), pp. 70–88.

Stephen, A., Bart, Y., and Plessis, C. Du. 2012. "Does Paying For Online Product Reviews Pay Off? The Effects of Monetary Incentives on Content Creators and Consumers," *NA-Advances in Consumer Research* (40), pp. 228–231.

Underwood, H., and Findlay, B. 2004. "Internet Relationships and Their Impact on Primary Relationships," *Behaviour Change* (21:02), pp. 127–140.

Wang, A., Zhang, M., and Hann, I. 2018. "Socially Nudged: A Quasi-Experimental Study of Friends' Social Influence in Online Product Ratings," *Information Systems Research* (Forthcoming).

Wang, Y., Goes, P., Wei, Z., and Zeng, D. 2017. "Production of Online Word-Of-Mouth: Peer

Effects and the Moderation of User Characteristics."

Wang, Z. 2010. "Anonymity, Social Image, and the Competition for Volunteers: A Case Study of the Online Market for Reviews," *The B.E. Journal of Economic Analysis & Policy* (10:1), pp. 1–34.

Wasko, M., and Faraj, S. 2005. "Why Should I Share? Examining Social Capital and Knowledge Contribution in Electronic Networks of Practice," *MIS Quarterly* (29:1), pp. 35–57.

Wasko, M., Teigland, R., and Faraj, S. 2009. "The Provision of Online Public Goods: Examining Social Structure in an Electronic Network of Practice," *Decision Support Systems* (47:3), pp. 254–265.

Willer, R. 2009. "Groups Reward Individual Sacrifice: The Status Solution to the Collective Action Problem," *American Sociological Review* (74:1), pp. 23–43.

Xia, M., Huang, Y., Duan, W., and Whinston, A. B. 2011. "To Continue Sharing or Not to Continue Sharing ? – An Empirical Analysis of User Decision in Peer-to-Peer Sharing Networks," *Information Systems Research* (23:1), pp. 1–13.

Yelp, I. 2011. "Yelp and the '1/9/90 Rule.'" (https://www.yelpblog.com/2011/06/yelp-and-the-1990-rule).

Yin, D., Bond, S., and Zhang, H. 2014. "Anxious or Angry? Effects of Discrete Emotions on the Perceived Helpfulness of Online Reviews," *MIS Quarterly* (38:2), pp. 539–560.

Zhang, J., Liu, Y., and Chen, Y. 2015. "Social Learning in Networks of Friends versus Strangers," *Marketing Science* (34:4), pp. 573–589.

Zhang, X., and Zhu, F. 2011. "Group Size and Incentives to Contribute : A Natural Experiment at Chinese Wikipedia," *American Economic Review* (101:June), pp. 1601–1615.

Adomavicius, G., Bockstedt, J. C., Curley, S. P., and Zhang, J. 2013. "Do recommender systems manipulate consumer preferences? A study of anchoring effects," Information Systems Research (24:4), pp. 956–975.

Anderson, E., and Sullivan, M. 1993. "The antecedents and consequences of customer satisfaction for firms," Marketing science (12:2), pp. 125–143.

Chevalier, J., and Mayzlin, D. 2006. "The effect of word of mouth on sales: Online book reviews," Journal of marketing research (43:3), pp. 345–354.

CNET. 2016. "Amazon continues crackdown on alleged fake reviews," https://www.cnet.com/news/amazon-continues-crack-down-on-alleged-fake-reviews-site.

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), pp. 127–158.

Dellarocas, C., and Narayan, R. 2006. "A statistical measure of a population's propensity to engage in post-purchase online word-of-mouth," Statistical Science (21:2), pp. 277–285.

Dellarocas, C., and Wood, C. 2008. "The sound of silence in online feedback: Estimating trading risks in the presence of reporting bias," Management Science (54:3), pp. 460–476.

Gao, G., Greenwood, B., McCullough, J., and Agarwal, R. 2012. "Noisy Minority and Silent Majority: an Examination of Biases in Online Physician Ratings," Working Paper (31:3), pp. 448–473.

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

Goes, P. O., Lin, M., and Yeung, C. A. 2014. "'Popularity Effect' in User-Generated Contents: Evidence from Online Product Reviews," Information Systems Research (25:2), pp. 222–238.

Goldberg, K., Roeder, T., Gupta, D., and Perkins, C. 2001. "Eigentaste: A constant time collaborative filtering algorithm," Information Retrieval (4:2), pp. 133–151.

Gu, B., Park, J., and Konana, P. 2012. “Research note-the impact of external word-of-mouth sources on retailer sales of high-involvement products,” Information Systems Research (23:1), pp. 182–196.

Higgins, E. 1996. “Knowledge activation: Accessibility, applicability, and salience,” in Social psychology: Handbook of basic principles, New York: The Guilford Press, pp. 133–168.

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

Ke, Z., and Liu, D. 2015. “Peer Effects and the Production of Online Reviews: A Message Level Analysis,” Thirty Sixth International Conference on Information Systems, Fort Worth 2015.

Lee, Y., Tan, Y., and Hosanagar, K. 2015. “Do Ifollow My Friends or the Crowd? Information Cascades in Online Movie Rating,” Management Science (61:9), pp. 2241–2258.

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

Liu, Y., Chen, P., and Hong, Y. 2014. “Matching Consumer Preferences with Product Attributes: The Value of Multi-dimensional Online Word of Mouth Systems,” Proceedings of ICIS.

Lu, X., Ba, S., Huang, L., and Feng, Y. 2013. “Promotional marketing or word-of-mouth? Evidence from online restaurant reviews,” Information Systems Research (24), pp. 596–612.

Luca, M. 2011. “Reviews, reputation, and revenue: The case of Yelp. com,” Com (September 16, 2011). Harvard Business School.

Ma, X., Khansa, L., Deng, Y., and Kim, S. S. 2013. “Impact of Prior Reviews on the Subsequent Review Process in Reputation Systems,” Journal of Management Information Systems (30:3), pp. 279–310.

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

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

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

Oliver, R. 1980. “A cognitive model of the antecedents and consequences of satisfaction decisions,” Journal of marketing research, pp. 460–469.

Schweidel, D., and Moe, W. 2014. “Listening in on social media: A joint model of sentiment and venue format choice,” Journal of Marketing Research (51:4), pp. 387–402.

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

Sridhar, S., and Srinivasan, R. 2012. “Social influence effects in online product ratings,” Journal of Marketing (76:5), pp. 70–88.

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

The New York Times. 2011. “A Rave, a Pan, or Just a Fake?,” http://www.nytimes.com/2011/05/22/your-money/22haggler.html?_r=0.

Tversky, A., and Kahneman, D. 1974. “Judgement under uncertainty Heuristics and biases,” Science (185), pp. 1124–1130.

Wang, A., Zhang, M., and Hann, I. 2015. “Socially Nudged: A Quasi-Experimental Study of Friends’ Social Influence in Online Product Ratings,” Information Systems Research, (Forthcoming).

Wang, Z. 2010. “Anonymity, Social Image, and the Competition for Volunteers: A Case Study of the Online Market for Reviews,” The B.E. Journal of Economic Analysis & Policy (10:1), pp.

1–34.

Wu, F., and Huberman, B. 2010. “Opinion formation under costly expression,” ACM Transactions on Intelligent Systems and Technology (TIST) (1:1), A. 5 (1-13).

Yelp Support Center. “How is the order of reviews determined?,” http://yelp-support.force.com/article/How-is-the-order-of-reviews-determined?l=en_GB.

Zeng, X., and Wei, L. 2013. “Social ties and user content generation: Evidence from Flickr,” Information Systems Research (24:1), pp. 71–87.

Anderson, E., and Simester, D. 2014. “Reviews without a Purchase: Low Ratings, Loyal Customers, and Deception,” Journal of Marketing Research (51:3), pp. 249–269.

Bacharach, S. B. 1989. “Organizational Theories: Some Criteria for Evaluation,” Academy of Management Review (14:4), pp. 496–515.

Ballou, D. P., and Pazer, H. L. 1985. “Modeling Data and Process Quality in Multi-Input, Multi-Output Information Systems,” Management Science (31:2), pp. 150–162.

Birdwhistell, R. 2010. Kinesics and Context: Essays on Body Motion Communication, University of Pennsylvania press.

BrightLocal. 2018. “Local Consumer Review Survey.” (https://www.brightlocal.com/research/local-consumer-review-survey/).

Broadbent, D. 1958. Perception and Communication, Oxford: Oxford University Press.

Cao, Q., Duan, W., and Gan, Q. 2011. “Exploring Determinants of Voting for the ‘Helpfulness’ of Online User Reviews: A Text Mining Approach,” Decision Support Systems (50:2), pp. 511–521.

Chen, C., and Tseng, Y. 2011. “Quality Evaluation of Product Reviews Using an Information Quality Framework,” Decision Support Systems (50:4), pp. 755–769.

Cheung, C., Lee, M., and Rabjhon, N. 2008. “The Impact of E-WOM — The Adoption of Online Opinions in Online Customer Communities.,” Internet Research (18:3), pp. 229–247.

Chou, C., Fatemeh" Mariam" Zahedi, and Zhao, H. 2014. “Ontology-Based Evaluation of Natural Disaster Management Websites: A Multistakeholder Perspective.,” MIS Quarterly (38:4), pp. 997–1016.

Clark, J., and Paivio, A. 1991. “Dual Coding Theory and Education,” Educational Psychology Review (3:3), pp. 149–210.

DeLone, W., and McLean, E. 1992. “Information Systems Success: The Quest for the Dependent Variable,” Information Systems Research (3:1), pp. 60–95.

Deutsch, M., and Gerard, H. B. 1955. “A Study of Normative and Informational Social Influences upon Individual Judgement.,” The Journal of Abnormal and Social Psychology (51:3), pp. 629–636.

Driver, J. 2001. “A Selective Review of Selective Attention Research from the Past Century,” British Journal of Psychology (92:1), pp. 53–78.

Eppler, M., and Wittig, D. 2000. “Conceptualizing Information Quality: A Review of Information Quality Frameworks from the Last Ten Years.,” in Proceedings of the 2000 Conference on Information Quality.

Eslami, S., Ghasemaghaei, M., Systems, K. H.-D. S., and 2018, U. 2018. “Which Online Reviews Do Consumers Find Most Helpful? A Multi-Method Investigation,” Decision Support Systems (113:Sep 1), pp. 32–42.

Forman, C., and Ghose, A. 2008. “Examining the Relationship between Reviews and Sales: The Role of Reviewer Identity Disclosure in Electronic Markets,” Information Systems Research (19:3), pp. 291–313.

Ghose, A., and Ipeirotis, P. 2011. “Estimating the Helpfulness and Economic Impact of Product

Reviews: Mining Text and Reviewer Characteristics," IEEE Transactions on Knowledge and Data Engineering (23:10), pp. 1498–1512.

Goodhue, D. L. 1995. "Understanding User Evaluations of Information Systems," Management Science (41:12), pp. 1827–1844.

Green, P. 1982. "The Content of a College-Level Outdoor Leadership Course.," in Paper Presented at the Conference of the Northwest District Association for the American Alliance for Health, Physical Education, Recreation, and Dance, Spokane, WA.

Gruber, T. 1995. "Toward Principles for the Design of Ontologies Used for Knowledge Sharing?," International Journal of Human-Computer Studies (43:5–6), pp. 907–928.

Hamilton, D., and Huffman, L. 1971. "Generality of Impression-Formation Processes for Evaluative and Nonevaluative Judgments.," Journal of Personality and Social Psychology (20:2), p. 200.

Hedlin, D., Dale, T., Haraldsen, G., and Jones, J. 2005. "Burden, Developing Methods for Assessing Perceived Response," Statistics Swden/Statistics Norway/UK Office for National Statistics.

Holsapple, C., and Joshi, K. 2002. "A Collaborative Approach to Ontology Design," Communications of the ACM (45:2), pp. 42–47.

Huang, K., Lee, Y., and Wang, R. 1999. Quality Information and Knowledge, New Jersey: Prentice Hall.

Jarke, M., and Vassiliou, Y. 1997. "Data Warehouse Quality: A Review of the DWQ Project.," in Proceedings of the Conference on Information Quality (Vol. 299–313), Cambridge, MA.

Kahn, B., and Strong, D. 1998. "Product and Service Performance Model for Information Quality: An Update," in Proceedings of the 1998 Conference on Information Quality, Cambridge, MA: Massachusetts Institute of Technology.

Kanfer, R., and Ackerman, P. 1989. "Motivation and Cognitive Abilities: An Integrative/Aptitude-Treatment Interaction Approach to Skill Acquisition.," Journal of Applied Psychology (74:4), p. 657.

Karimi, S., Wang, F., and 2017, undefined. 2017. "Online Review Helpfulness: Impact of Reviewer Profile Image," Decision Support Systems (96), pp. 39–48.

Kerlinger, F. 1986. Foundations of Behavioral Research., New York: Harcourt Brace Jovanovich College Publishers.

Koehler, M., Yadav, A., and Phillips, M. 2005. "What Is Video Good for? Examining How Media and Story Genre Interact," Journal of Educational Multimedia and Hypermedia (14:3), pp. 249–272.

Lee, Y. W., Strong, D. M., Kahn, B. . K., and Wang, R. Y. 2002. "AIQM: A Methodology for Information Quality Assessment," Information & Management (40:2), pp. 133–146.

Lesca, H., Lesca, E., Lesca, N., and Caron-Fasan, M. 1995. Gestion de l'information: Qualité de l'information et Performances de l'entreprise, Paris: Litec.

Linton, H., Han, S., and Gretzel, U. 2017. "TripAdvisor Super Contributors: Projecting Professionalism," in Frontiers in Service Conference, New York.

Liu, J., Cao, Y., Lin, C., Huang, Y., and Zhou, M. 2007. "Low-Quality Product Review Detection in Opinion Summarization," in Computational Linguistics, pp. 334–342.

Lu, Y., Tsaparas, P., Ntoulas, A., and Polanyi, L. 2010. "Exploiting Social Context for Review Quality Prediction," in Proceedings of the 19th International Conference on World Wide Web, ACM, pp. 691–700.

Malik, M., and Hussain, A. 2017. "Helpfulness of Product Reviews as a Function of Discrete Positive and Negative Emotions," Computers in Human Behavior (73), pp. 290–302.

Muchnik, L., Aral, S., and Taylor, S. 2013. “Social Influence Bias: A Randomized Experiment,” Science (341:6146), pp. 647–651.

Mudambi, S., and Schuff, D. 2010. “What Makes a Helpful Review? A Study of Customer Reviews on Amazon. Com,” MIS Quarterly, pp. 185–200.

Okoli, C., and Pawlowski, S. 2004. “The Delphi Method as a Research Tool: An Example, Design Considerations and Applications,” Information & Management (42:1), pp. 15–29.

Paivio, A. 1990. Mental Representations: A Dual Coding Approach, (9th ed.), Oxford University Press.

Pan, Y., and Zhang, J. Q. 2011. “Born Unequal: A Study of the Helpfulness of User-Generated Product Reviews,” Journal of Retailing (87:4), pp. 598–612.

Park, D., Lee, J., and Han, I. 2007. “The Effect of Online Consumer Reviews on Consumer Purchasing Intention: The Moderating Role of Involvement,” International Journal of Electronic Commerce (11:4), pp. 125–148.

Pavlou, P., Liang, H., and Xue, Y. 2007. “Understanding and Mitigating Uncertainty in Online Environments: A Principal-Agent Perspective,” MIS Quarterly (31:1), pp. 105–136.

Pribram, K., and McGuinness, D. 1975. “Arousal, Activation, and Effort in the Control of Attention.,” Psychological Review (82:2), p. 116.

Sadoski, M., Goetz, E., and Rodriguez, M. 2000. “Engaging Texts: Effects of Concreteness on Comprehensibility, Interest, and Recall in Four Text Types.,” Journal of Educational Psychology (92:1), p. 85.

Salehan, M., and Kim, D. 2016. “Predicting the Performance of Online Consumer Reviews: A Sentiment Mining Approach to Big Data Analytics,” Decision Support Systems (81:Jan 1), pp. 30–40.

Salisbury, C., Burgess, A., Lattimer, V., and Heaney, D. 2005. “Developing a Standard Short Questionnaire for the Assessment of Patient Satisfaction with Out-of-Hours Primary Care,” Family Practice (22:5), pp. 560–569.

Sternberg, R. . 2003. Cognitive Theory, Thomson Wadsworth, Belmont.

Taylor, S. 1991. “Asymmetrical Effects of Positive and Negative Events: The Mobilization-Minimization Hypothesis.,” Psychological Bulletin (110:1), p. 67.

Treisman, A. 1969. “Strategies and Models of Selective Attention.,” Psychological Review (76:3), p. 282.

Ullah, R., Zeb, A., and Kim, W. 2015. “The Impact of Emotions on the Helpfulness of Movie Reviews,” Journal of Applied Research and Technology (13:3), pp. 359–363.

Wand, Y., and Wang, R. 1996. “Anchoring Data Quality Dimensions in Ontological Foundations,” Communications of the ACM (39:11), pp. 86–95.

Wang, R., and Strong, D. 1996. “Beyond Accuracy: What Data Quality Means to Data Consumers,” Journal of Management Information Systems (12:4), pp. 5–33.

Yechiam, E., and Hochman, G. 2013. “Losses as Modulators of Attention: Review and Analysis of the Unique Effects of Losses over Gains.,” Psychological Bulletin (139:2), p. 497.

Yin, D., Bond, S., and Zhang, H. 2014. “Anxious or Angry? Effects of Discrete Emotions on the Perceived Helpfulness of Online Reviews,” MIS Quarterly (38:2), pp. 539–560.

Zmud, R. 1978. “An Empirical Investigation of the Dimensionality of the Concept of Information,” Decision Sciences (9:2), pp. 187–195.

Appendices

Appendix A

Table 4.1. Ratings on the Anchoring Definition

rating	clarity		correctness		completeness		conciseness	
	#	%	#	%	#	%	#	%
1	1	6	1	6	1	6	1	6
2	0	0	1	6	2	11	0	0
3	6	33	1	6	4	22	4	22
4	9	50	12	67	9	50	6	33
5	2	11	3	17	2	11	7	39

# (%): the number (percent) of participants give the corresponding rating in the row

Table 4.2. Ratings on the Completeness of the Anchoring Dimension List

rating	#	%
1	0	0
2	4	22
3	4	22
4	9	50
5	1	6

# (%): the number (percent) of participants give the corresponding rating in the row

Table 4.3. Ratings on the Pertinency of the Anchoring Individual Dimensions

rating	relevant	trust*	objective	author*	original	compr*	organized	concise	infor*	enter*	timely
	%	%	%	%	%	%	%	%	%	%	%
1	0	6	6	11	0	0	0	6	0	11	0
2	0	6	6	17	22	0	6	0	0	50	0
3	0	6	11	22	22	0	22	44	6	33	17
4	6	22	33	28	50	22	56	33	11	6	39
5	94	61	44	22	6	78	17	17	83	0	44

* trust: trustworthy; author: authoritative; compr: comprehensible; info: informative; enter: entertaining.
%: the percentage of participants give the corresponding rating in the row

Appendix B

Table 4.4. Ratings on the Revised Definition

rating	clarity		correctness		completeness		conciseness	
	#	%	#	%	#	%	#	%
1	0	0	1	10	1	10	0	0
2	0	0	0	0	0	0	0	0
3	1	10	1	10	0	0	0	0
4	4	40	3	30	4	40	5	50
5	5	50	5	50	5	50	5	50

# (%): the number (percent) of participants give the corresponding rating in the row

Table 4.5. Ratings on the Completeness of the Revised Dimension List

rating	#	%
1	0	0
2	0	0
3	0	0
4	4	40
5	6	60

# (%): the number (percent) of participants give the corresponding rating in the row

Table 4.6. Ratings on the Pertinency of the Revised Individual Dimensions

rating	relevant		trustworthy		comprehensive		well-written		timely	
	#	%	#	%	#	%	#	%	#	%
1	0	0	0	0	0	0	0	0	0	0
2	0	0	1	0	0	0	0	0	0	0
3	3	30	0	0	0	0	0	0	1	10
4	3	30	2	20	1	10	3	30	1	10
5	4	40	7	70	9	90	7	70	8	80

# (%): the number (percent) of participants give the corresponding rating in the row