Published in final edited form as:

Decis Support Syst. 2020 January ; 128: . doi:10.1016/j.dss.2019.113158.

A Matter of Reevaluation: Incentivizing Users to Contribute Reviews in Online Platforms

Mingyue Zhang^a, Xuan Wei^{b,c,*}, Daniel Dajun Zeng^{d,e}

^aSchool of Business and Management, Shanghai International Studies University, Shanghai 201620, China

^bEller College of Management, University of Arizona, Tucson, AZ 85721, United States

^cAntai College of Economics and Management, Shanghai Jiao Tong University, Shanghai 200030, China

^dInstitute of Automation, Chinese Academy of Sciences, Beijing 100864, China

^eShenzhen Artificial Intelligence and Data Science Institute (Longhua), Shenzhen 518129, China

Abstract

Content sharing platforms such as product review websites largely depend on reviewers' voluntary contributions. In order to motivate reviewers to contribute more, many platforms established incentive mechanisms, either reputation-based or financial. Yet most of the existing research has focused on reputations that are everlasting, such as badges and virtual points, or financial rewards where no evaluation exists about the users' contributed content, such as rebates. There is still a significant gap in our understanding of how incentives with reevaluation mechanism actually influence reviewers' behaviors such as their contribution levels, the opinion they express, and how they express. In this paper, we fill this gap using data collected from Yelp Elite Squad where reviewers with good reviewing history are awarded into the elite group and most importantly reevaluated each year. We draw from the accountability theory and conduct a difference-in-differences analysis to empirically study the effect of incentives with reevaluation mechanism on reviewers' behaviors in both short term and long term. The results show that in short term, reviewers significantly increase their contribution levels, become more conservative with lower percentage of extreme ratings, and also increase the readability of their reviews. In long term, they continue improving the quality of reviews though their numerical rating behaviors stabilize. Our research has significant implications for business models that rely on user contributions.

Keywords

Reevaluation mechanism; incentives; accountability; product reviews

*Corresponding author weix@email.arizona.edu (Xuan Wei).

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

1. Introduction

Nowadays, an increasing number of websites such as e-commerce sites, social network sites, blogs, and video, and image sharing sites, heavily rely on user-generated content (UGC). Among these content sharing platforms, product reviews sites are one of the most common platforms. Salient examples include Yelp, TripAdvisor, and RottenTomatoes, all of which are non-commercial review sites and have become more and more popular in the past few decades [1]. There is mounting evidence indicating that such online reviews significantly influence consumer choices and product sales [2, 3, 4, 5, 6, 7]. The voluntary contributions from the community are the core power of generating online reviews, retaining existing users, as well as attracting new ones [8]. Given the importance of user contribution, a million-dollar question is why users devote their valuable time and effort to voluntarily contribute new contents and help strangers in UGC sites?

Actually, the voluntary contribution does not always happen naturally in that there exists potential free riding phenomenon in UGC sites [9]; it is commonly documented that a small proportion of users often account for the majority of the content according to Pareto principle [10]. To motivate the users to contribute more as well as higher quality contents, the UGC sites have built various incentive mechanisms, aiming at increasing the benefits from users' efforts besides their intrinsic motivations [11]. For example, [Amazon.com](https://www.amazon.com) launched the Amazon Vine Program in 2007, through which reviewers write honest and unbiased product reviews in exchange of the eligibility to receive free or discounted sample products [12]. EBay users could receive rebates by providing feedback after purchasing [13]. In addition to financial rewards, reputation-based incentives are also widely deployed. For instance, by answering questions in a knowledge sharing platform, users may receive virtual points and thus reach higher rank in an incentive hierarchy among their peers [14]. Some other sites provide different badges for various aspects of a user's activities [15, 8].

The proliferation of incentive mechanisms has drawn much attention from the academic community where they attempt to explore whether the incentive mechanisms are effective in inducing users to contribute more and higher quality contents. Though various incentive types (i.e., financial or reputation-based) have been intensively studied, most existing research focuses on the scenario where no reevaluation of users' performance exists. For reputation-based incentives such as badges¹ and points, they are often everlasting [15, 14]. The reviewers or other type of content producers can always keep the glory badges or accumulated points once acquiring them, even though they may stop contributing. When the incentives are financial rewards such as rebates, existing literature does not consider the platforms where users' performance is evaluated [16, 17]. For example, sellers provide a rebate to cover the buyer's reporting cost, regardless of whether the feedback is positive or negative and how long and informative the feedback is [16]. The *reevaluation mechanism*, as a potential means to incentivize users to contribute more and higher quality contents, is largely under-explored. Particularly, *reevaluation mechanism* means users need to be reevaluated periodically based on their contributing performance to maintain the extrinsic

¹Hereafter, we reserve the term "badge" for traditional badge systems where badges are everlasting, although Yelp also uses some signs—badges in a broader sense—to signal the advanced status.

benefits—either financial or reputation-based. Under such a mechanism, users bear the pressure of losing the achieved reputation or financial perks from previous period. This is quite common in UGC sites such as Yelp. Yelp Elite Squad program was launched in 2005, and serves as one way of recognizing reviewers who are active in the Yelp community and role models on and off the site. The elite reviewers can attain higher reputation compared with their peers and enjoy some associated perks. The program is a yearly program and reviewers who were admitted into the Elite Squad need to be reevaluated every year.

In this work, we focus on the scenario where incentives are implemented along with reevaluation mechanism. This is essentially different from the traditional badges, points accumulating programs, or financial incentives (hereafter, “traditional settings”). Intuitively, under reevaluation settings, users need to continue contributing more and higher quality contents in order to maintain their acquired benefits, while users bear no pressure of losing the benefits in traditional settings. Note that the incentives with reevaluation mechanism can be either financial rewards or non-financial benefits such as higher reputation. This means the existence of reevaluation is not limiting the scope of our research; on the contrary, it applies to various incentives including financial rewards or non-financial benefits. Given this widespread, important, and under-researched research scenario, it is urgent and necessary that we systematically examine whether and how the incentives with reevaluation mechanism influence user behaviors. If there is any behavioral change, it’s also important to study whether the change is temporary or long-lasting.

To fill the observed research gap, we study how reviewers in the Yelp Elite Squad change their behaviors after they receive the incentives with reevaluation mechanism (i.e., being Yelp elites) in both short term and long term. In our research context, reviewers are former consumers of the business who write comments based on their own experience and perception. Their reviews may serve as a key information source about product quality that significantly affects other consumers’ choices and product sales [2, 3]. Among all reviewers, those role models on and off the site are awarded into the “Elite Squad”. The “Elite” signs on their Yelp profiles will demonstrate a higher reputation among peers. In addition, being elite also comes with financial perks such as exclusive invitations to private parties held in local businesses [18]. Most importantly, elite reviewers need to be reevaluated every calendar year based on their contributing behaviors in order to stay in the “Elite Squad”. To clarify, we are not trying to separate the financial incentive effect from the reputation incentive effect in this research context; instead, we focus on reviewers’ behavioral changes when there is reevaluation for them to maintain the benefits (either financial or non-financial benefits). Hence, we propose the following research questions:

How does the incentives with reevaluation mechanism influence reviewer behaviors, including (1) the contribution level (frequency of reviewing and length of reviews), (2) opinions that they express (average rating, variance, and extreme rating), and (3) how they express (the quality of reviews)? If there is any influence, is it temporary or long lasting?

Answering these research questions is challenging. For example, there are two potential counterarguments for the contribution level. On one hand, being recruited into Elite Squad may be a goal for reviewers, and hence, their contribution levels may drop significantly after

reaching the goal (i.e., being elites) due to “complacency” effects [19, 14]. On the other hand, the Elite sign prominently displays next to elite reviewers’ names on the reviewers’ profile avatars, as well as all reviews that they wrote. As a result, it reveals and signifies the elite reviewers’ reputation in the community and distinguishes them from their peers. Drawing from prospect theory [20], reviewers are afraid of being eliminated from the Elite Squad next year since it is readily observed by others. Therefore, another possible change in behavior is that they review in higher quality and frequency to maintain the reputation as well as the associated perks, which in some sense, is consistent with previous literature indicating that reputation and recognition motivate people to behave in a better way [19, 21].

To systematically answer the proposed questions, we collect the profiles of elite reviewers in Yelp and analyze their behavioral changes from the year before being elites to the first year of being elites (i.e., short-term effect), and then to the second year of being elites (i.e., long-term effect). Additionally, to account for potential endogeneity issues due to reviewers’ self-selection, we collect the profiles of non-elite reviewers and combine a propensity score matching (PSM) and a difference-in-differences (DID) approach to conduct further empirical analysis. This enables us to simulate a quasi-experimental environment and estimate the “treatment effect” of incentives with reevaluation mechanism on reviewers’ behaviors.

The rest of this paper is organized as follows. We review related literature and develop our hypotheses in Section 2 and Section 3, respectively. Details of research design including data collection and model specification are introduced in Section 4. Section 5 describes the estimation results, as well as some robustness checks. Section 6 discusses contributions and concludes the paper.

2. Related Literature

2.1. Incentive Mechanism

Our study mainly builds upon and contributes to the research of incentive mechanism on UGC sites. To motivate users to contribute more contents, UGC sites often build various incentive mechanisms to increase the users’ extrinsic motivations [22]. The incentives might be financial rewards such as free gifts [12], rebates [13], and extra payments [23], or reputation-based rewards such as higher social status [14, 23, 24], social comparison [9, 25], and achievement badges [15, 8]. For reputation-based incentives, it is worth noting that reputation systems are widely used by most electronic markets to mitigate problems arising from information asymmetry, and the design varies across websites depending on whether the stakeholders are sellers, buyers or reviewers [26, 27]. In online exchange markets like eBay, reputation systems are built for sellers and buyers to facilitate trustworthy transactions among strangers through feedback mechanism [28, 29, 30]. In professional review sites like Yelp or Epinions, the reputation of reviewers can be measured by their number of prior reviews or number of friends/followers [1, 19]. Some sites build reputation systems for the reviewers by allowing them to receive points when making contributions or granting them various accolades or certificates to recognize them in front of their peers [14]. Without loss of generality, in this paper we refer to the reputations for *reviewers* when using the term “reputation”.

Evidence from prior literature shows that users' contribution behaviors are strongly influenced by incentive mechanisms in various contexts [31, 25]. (1) In open source software communities, Roberts et al. (2006) [23] studied the effects of both financial and non-financial incentives on users' contributions based on theories of intrinsic and extrinsic motivation. They found that being paid to contribute is positively related to users' status motivations but negatively related to their use-value motivations, and consequently, this leads to above-average contribution levels. (2) In online Q&A communities or knowledge exchange communities, Goes et al. (2016) [14] drew on goal setting and status hierarchy theories to study users' contributions before and after they reach consecutive ranks on a vertical incentive hierarchy. Another example is Wasko and Faraj's study (2005) [24] where they found that users contributed their knowledge when they perceived that it enhanced their professional reputations according to the theory of collective action. Similarly, Li et al. (2012) [8] identified a short-term positive effect of winning new badges in Q&A communities. (3) In product review sites or e-commerce sites where review is an important component, Goes et al. (2014) [19] examined the effect of reviewers' popularity on their review behaviors under subscription mechanism where one reviewer is allowed to subscribe to another. They found that such a mechanism was effective in inducing reviewer efforts, that is, reviewers produce more reviews and more objective reviews as they become more popular. Using monetary rebates as incentives for feedback, Li and Xiao (2014) [17] conducted a laboratory experiment to examine the effect of the rebate incentive on market efficiency in a listed-price e-market. They found that market efficiency under the rebate mechanism increases with the probability that sellers will provide a rebate. Qiao et al. (2017) [32] argued that monetary incentive provision in review platforms would greatly damage the reviewers' original altruistic and intrinsic motivations and resulted in lower quality and less helpful reviews in short term.

In summary, prior work has discussed several consequences of financial incentives or reputation-based incentives. Given that there is virtually no empirical evidence showing the effectiveness of incentives when reevaluation about users' performance exists, our study aims to fill this gap by examining how incentives with reevaluation mechanism affect the users' behaviors, regardless of the incentive types. In our research context, elite reviewers in Yelp need to be reevaluated every year to maintain their higher reputation (i.e., elite) as well as the associated perks. It's very likely that findings in existing studies do not apply to our scenario. It is expected that reviewers behave differently under the reevaluation mechanism and our research will contribute to the growing literature by illuminating new insights on the effect of incentives with reevaluation mechanism.

2.2. Accountability Theory

In this subsection, we will briefly review relevant literature that lays the theoretical foundation for the proposed hypotheses. The accountability theory was originally developed by Lerner and Tetlock (1999) [33], and then widely applied in a variety of fields, including psychology, philosophy, ethics, and organizational behavior [34]. It is a process in which a person has a potential obligation to explain her/his activities toward another party who can make judgment on these activities and also administer potential positive or negative consequences as a response to them [35]. The theory also demonstrates two key elements

that stimulate accountability perceptions: overt expectations of evaluation and awareness of monitoring. Later, Vance et al. [36, 34] extended the accountability theory to Information Systems (IS) context and developed four user-interface design artifacts to raise users' accountability perceptions within systems. Specifically, four constructs, namely, identifiability, expectation of evaluation, awareness of monitoring, and social presence will raise accountability. Identifiability refers to a person's "knowledge that her/his outputs could be linked to her/him" and thus reveals her/his true identity [37]. Therefore, individuals who perceive increased identifiability know that they can be made responsible for their actions [33]. Second, expectation of evaluation is the belief that one's activities will be assessed and judged by others with some implied consequences [33]. Such awareness will cause socially desirable behaviors [38]. Third, monitoring is the process of tracking one's activities. Awareness of monitoring will increase the user's expectation that s/he is accountable. Finally, individuals exhibit increased conforming behavior when they are aware of other users in the system, namely, social presence. Based on the theory, extensive literature also studied the outcomes of increased perception of accountability. Particularly, users who perceive themselves to be accountable to the systems are more likely to achieve a cognitive awareness that will increase pro-social behaviors [39], increase conformity to expected behaviors [40], increase conservatism [41], and decrease risk taking [42].

3. Hypotheses Development

In this section, we develop the main hypotheses from three aspects: (1) review volume and length which reveal reviewers' contribution levels; (2) numerical ratings of reviews which reflect the opinions that the reviewers express; and (3) readability of reviews which is an important indicator of how the opinions are expressed.

The reevaluation mechanism on Yelp Elite Squad program will create a sense of accountability, thus influencing elite reviewers' contribution levels. Specifically, the elite reviewers have a clear expectation of evaluation if they want to be elites in the coming calendar year. Such an expectation is an important component of accountability [36, 34]. In addition, elite reviewers are required to disclose their real names and post real photos in the profile from which they experience more identifiability and individuation [34]. This fact further makes them feel they are being monitored, especially when they are aware of the yearly reevaluation. These two factors (i.e., identifiability and awareness of monitoring) will both increase the sense of accountability [35]. Last, members of Elite Squad are also more likely to be recognized than the average reviewers in the virtual community, as well as occupying a structurally advantageous position within the social network [43]. As a result, reviewers have more exposure chances (i.e., higher social presence) to the public after they become elites. In summary, the reevaluation mechanism on Yelp Elite Squad program increases elite reviewers' perceived accountability towards the system. According to the accountability theory [34], elite reviewers with the sense of accountability will behave in socially desirable manners. It's natural to expect that elite reviewers contribute more and longer reviews comparing to their peers.

Another reason for the elite reviewers to behave in certain socially desirable manners is that they want to maintain the benefits of being elites. The Yelp Elite Squad program is

associated with financial incentives (e.g., perks²) for elite reviewers in addition to the high reputation. For example, elite reviewers may be invited to private parties held in local businesses. For some other elite reviewers, the associated high reputation may be the driving force of becoming elites, consistent with the widely recognized reputation-seeking behaviors on online review websites [24, 44]. Due to the existence of the reevaluation mechanism, existing elite reviewers need to behave in certain ways to maintain the acquired elite reputation as well as other benefits, especially by writing more and longer reviews, which are the most straightforward measures of contribution levels.

For the long-term effect, the accountability and motivation to maintain the benefits should be stronger at the beginning when they were first admitted into this Squad [19]. Using review length as the indicator of contribution level, the increasing rate should decrease until it reaches a stable state. The number of reviews will decrease in long term because there may be certain “complacency” effect after the reviewers passing the first reevaluation process (i.e., the second year after being elite). We therefore hypothesize the following:

H1: (a) After receiving incentives with reevaluation mechanism, reviewers will increase their contribution levels such as the number of reviews and length of reviews in short term. (b) In long term, the number of reviews will decrease while the length remains stable.

The opinions reviewers express may also be influenced by becoming elites, which can be reflected by the numerical ratings. Here we focus on three aspects of ratings: the average rating, variance of ratings, and ratio of extreme ratings. As mentioned earlier, elite reviewers that bear the pressure of reevaluation will experience an expectation of evaluation, as well as higher identifiability, awareness of monitoring, and social presence compared with those non-elite reviewers. This will elicit an increased sense of accountability [34]. With accountability towards the information system, reviewers will increase conservatism [41] and decrease risk taking [42]. Thus, elite reviewers will give fewer negative ratings to mitigate risks, resulting in higher average rating. For rating extremity, which means the degree of deviation from the average rating of all reviews [45, 46, 47, 43], we focus on the ratio of extreme negative ratings (i.e., one star) in main analysis³. The major reason is that most online product reviews follow an asymmetric bimodal distribution with *J* shape and hence the five-star ratings are not normally considered as extreme ratings. Previous research found that moderate messages could enhance source credibility [48] while rating extremity diluted the influence of reviewer credibility [43]. Thus, elite reviewers will write fewer reviews with extreme ratings to improve credibility, behaving in the way that the Yelp platform expects. Given that rating variance is positively related with the ratio of extreme reviews, it will also decrease after one reviewer becoming elite. In long term, elite reviewers still possess high accountability due to the existence of reevaluation requirements, which differentiates with the goal-driven process in traditional everlasting badge system [14]. Thus, reviewers will keep all numerical rating behaviors to mitigate the risk. The marginal effects of numerical ratings will decrease until stable. We put forward the hypothesis:

²<https://www.birdsof.fire.com/yelp-elite-perks-fancy-free-food-cool-experiences/>

³We use the ratio of five stars in robustness check.

H2: (a) After receiving incentives with reevaluation mechanism, reviewers will write reviews with significantly different opinions in short term, including higher average ratings, lower rating variance, and lower percentage of extreme ratings. (b) In long term, these numerical rating behaviors remain stable.

To measure the quality of reviews, a natural metric is the readability of review contents [19]. Reviewers who receive recognition of Elite Squad may be given a sign that signals a type of connoisseurship or expertise, which could fulfill one's self-enhancement need [44]. Elite reviewers also receive some associated perks, such as invited local parties. Given that the qualification to obtain such benefits needs to be reevaluated each calendar year, the elite reviewers tend to review in higher quality to maintain the benefits. The increased perception of accountability also elicits reviewers' conformity to expected behaviors [40], thus writing reviews with higher quality in short term. Similar to the numerical rating behaviors, the marginal effect of readability will decrease until stable. Thus, we hypothesize:

H3: (a) After receiving incentives with reevaluation mechanism, reviewers will write reviews with higher readability in short term. (b) In long term, the readability remains stable.

4. Research Design

4.1. Data Collection

To empirically test the above hypotheses, we collected the reviewers' information from Yelp⁴. The data was collected in October 2018. Since it is infeasible to retrieve all reviewers' information across the website, we focus on these from certain cities, namely, Phoenix and Tucson in United States. Concretely, a snowball crawling strategy was adopted. For each city, we started from the community manager and collected her/his first-degree friends who were also in the same city. Particularly, community managers are Yelp employees that live in the reviewers' metro and help the community get the most out of Yelp by answering questions, hosting events and getting folks involved. Many reviewers established friend relations with their corresponding community managers to get the latest news and seek help, especially the elite reviewers. Thus, it is reasonable to collect data by starting from the community manager. Then the similar process was repeated for the collected friends until we got the six-degree friends. Figure 1 visualizes the number of elite and non-elite reviewers as a function of *degree* in Phoenix and Tucson, respectively. We can observe that most elite reviewers are the first and second-degree friends of the community manager and there are no more new reviewers in the network after six degrees. Hence, it's safe to claim that we have collected almost all reviewers, including both elites and non-elites, in these two cities⁵. At the end, we got 963 elite and 118,314 non-elite reviewers in Phoenix, and 572 elite and 53,392 non-elite reviewers in Tucson. It is worth noting that the elite reviewers are just the minority group in Yelp—only less than 2% of the reviewers are in Elite Squad in our collected data sets⁶.

⁴<https://www.yelp.com>

⁵For those isolated reviewers that are not connected to anyone, excluding them has no influence on our analyses.

⁶The actual percentage in the whole platform across all cities is far less than 2%.

For each reviewer, we collected the year(s) s/he was elite (i.e, if s/he is an elite member), the registration time, number of friends, number of followers, hometown, and most importantly all review information including numerical rating, review text, review date, and the number of received votes (i.e., useful, funny, and cool). We also recorded its “distance” to community manager and named it as *degree*. *Degree n* means this reviewer is the n-th degree friend of the community manager. We introduce such a variable because exposure to more reviewers may change reviewers’ behavior due to the peer influence [49], making *degree* a necessary control variable. As a result, for the Phoenix dataset, we got 546, 505 reviews in total with 155, 995 written by elite reviewers and 390, 510 written by non-elite reviewers. For the Tucson dataset, we got 182, 064 reviews in total with 67, 276 written by elite reviewers and 114, 788 written by non-elite reviewers. Based on review content, we also derived two textual features. First, we calculated the review length by directly counting the number of words. To measure the readability of reviews, we use a widely-used metric: the Lexical Density (LD). This metric is derived from prior studies [19, 50, 51] and it is inversely correlated with readability. The calculation formula is as follows.

$$\text{Lexical Density(LD)} = \left(\frac{\text{Number of Unique Words}}{\text{Number of Words}} \right) \times 100 \% \quad (1)$$

LD measures the degree of information contained in texts. Higher density suggests that a text contains more information and is more difficult to read. It is worth noting that “readability” is a conceptual construct while LD is the operationalization or measurement of readability. It is quite common to state the relationship between constructs in hypotheses and use their operationalization in empirical analyses.

The descriptive statistics of the collected data are shown in Table 1. In following sections, we use Phoenix dataset to conduct main analyses and Tucson dataset as a robustness check.

4.2. Exploratory Analysis

We first introduce how we construct the dataset and operationalize the short-term and long-term effects. Based on the observation that Yelp Elite Squad program is a yearly program, we constructed a four-year-long panel dataset for the elite reviewers. Specifically, elite reviewers do not necessarily enter the elite program at the same year. Thus, the dataset was constructed by aligning the year that the reviewer became a member of “Elite Squad”, which contains information of two years before and two years after the elite status acquirement. In addition, we operationalize one year after being elite as short term and the second year after being elite as long term⁷.

We conducted some exploratory analyses to show model-free evidence as to how reviewers’ behaviors change after they become elites. Specifically, we set one month as the time unit and plotted their behavioral changes over the 48 months in which they became elite members in the 24th month. In accordance with our hypotheses, we developed following dependent variables and plotted their dynamics in Figure 2: *number of reviews*; *average*

⁷Limited by the dataset, we are unable to define short-term and long-term effects in finer granularity.

review length, *average rating*, *variance of ratings*, *ratio of one-star ratings*, and *readability (LD)*. Since it may not be meaningful to calculate the mean and standard deviation of ratings when there are only very few ratings for the reviewer, we filtered out reviewers with less than three ratings to calculate average rating and variance, as shown in Figures 2(c) and 2(d). Similar data processing is performed in following main analysis. We also examined the robustness of results using different minimum numbers of reviews to calculate mean and standard deviation and we reached similar conclusions. The figures clearly demonstrate that those variables roughly change in line with our hypotheses. For example, the *number of reviews* increases significantly after being elite (the 24th month) in short term and then drops in long term. Another indicator of contribution level, namely, *average review length*, increases immediately and then remains stable in the second year. Regarding the opinions expressed, the average rating increases while the variance and ratio of one star decrease significantly in short term. In long term, these numerical ratings remain at the same level. Last, the readability of reviews increases after one year being elite and then remains stable in the second year. Moreover, the figures suggest that reviewers' behaviors are quite unstable when they are regular reviewers of the platform—these dependent variables go up and down frequently in the first 24 months. However, their behaviors converge and stabilize after entering the “Elite Squad”.

4.3. Matching

While the earlier exploratory analyses reveal that reviewers change their behaviors after entering Elite Squad, it might be argued that this is attributed to inherent reviewers characteristic rather than a result of acquiring benefits from the Elite Squad. That is to say, reviewers who are more intrinsically motivated tend to nominate themselves into Elite Squad. To address this potential self-selection issue, we propose to implement propensity score matching method [52, 53]. To this end, we need to construct a “control group” with reviewers who have never been recruited into Elite Squad but have similar review behaviors to those in the “treatment group”. Some unique features of our context and data help assure the validity of such a method. First, reverse causality and simultaneity bias are mitigated in the panel data because we match the reviewers using activities before the “treatment start time” only. Second, since the Elite Squad program uses a nomination mechanism, it is possible to find a control group because not all reviewers enter Elite Squad automatically after achieving certain contribution levels or having certain behaviors.

We first describe how we conducted matching to test the short-term effect. Since short term is defined as one year after being elite and we need one year before being elite for matching purpose, we filtered the elite reviewers to keep those having review records for at least one year before and after being elite as the treatment group. As mentioned above, one challenge for matching is that elite reviewers do not necessarily enter the elite program at the same time. Hence, we aligned these elite reviewers by the year when they became a member of “Elite Squad”. We then aggregated reviewing information of each reviewer for one year before being elite ($t = 0$) and the year after being elite ($t = 1$) respectively. For each treated reviewers (elite reviewers), we need to match her/him with a control reviewers (non-elite reviewers) based on the one-year review history prior to the “treatment start time” (being elite). To construct a set of candidate control, we filtered out the non-elite reviewers who

have less than two years of review records—first year for matching and the second year for comparison to test the short-term effect in following steps. Similar to the treatment group, we aggregated the review information in each year for each control candidate. Then, we performed a Nearest Neighboring matching after using standard probit function to model each reviewer's probability of entering the Elite Squad. The year-level characteristics used for matching include *average review length*, *review volume*, *average numerical rating*, *variance*, *ratio of extreme ratings*, *average LD*, *average number of votes*, and *number of friends*. Once a pair matched, the “treatment start time” can be defined for reviewers in the control group; and her/his reviewing information in the following year was then reserved for following difference-in-differences estimation. Finally, 721 pairs were matched (Phoenix dataset) and a two-year-long panel dataset for both treatment group and control group before and after treatment time was constructed. To ensure that our matching is successful, we plotted the distribution of propensity score before and after matching between two groups in Figure 3. We can see the propensity score distribution of the control group after matching is almost identical to that of the treatment group, which suggests that matching is satisfactory and convincing. We further conducted statistical tests and concluded that the distributions of all variables are not significantly different between control and treatment group after each matching. See the results in Table 2.

To test the long-term effect, we conducted similar matching procedures except for a few distinctions. For treatment group, the reviewers need to have at least two years reviewing history after being elite since we define the second year after being elite as long term. When conducting matching, we only used the aggregated information of the year before being elite. For control group, we filtered out the non-elite reviewers with less than three years of reviewing history. This is because the first year would be used for matching and the last two years would be used for further examining the short-term and long-term effects, respectively. As a result, 485 pairs were matched (Phoenix dataset), and we obtained a three-year-long panel dataset for both the treatment and control group before and after treatment time. Similar statistical test has been conducted to show there is no significant differences between variables after matching. Due to space constraints, the details are omitted. Note that the reason for performing two matchings is to retain data observations as many as possible.

4.4. Difference-in-Differences Estimation

4.4.1. Short-term effect—Once matched pairs were identified, we extracted their activities before (*Status* = 0) and after (*Status* = 1) the “being elite” event occurs and then compared them in a difference-in-differences manner to account for potential confounding factors such as time trends and maturation in a one-group pre–post design [54]. As mentioned above, we are interested in reviewers' behavior change in following aspects: contribution level, opinions expressed, and how to express these opinions. Therefore, we developed six dependent variables: (1) *number of reviews*, (2) *average review length*, (3) *average rating*, (4) *variance of ratings*, (5) *ratio of one-star ratings*, and (6) *readability: Lexical Density*. To rule out some potential exogenous factors, we also chose some control variables, including degree, tenure, number of friends, number of followers, and whether the reviewers is in hometown or not. To estimate the short-term effect, we used the matched

two-year-long panel dataset such that each unit of observation is a reviewer and each time period is one calendar year, and specified our DID model as follows:

$$DV_{it} = \alpha_0 + \alpha_1 Treat_{it} + \alpha_2 Status_{it} + \alpha_3 Treat_{it} \times Status_{it} + \alpha_4 Cov_{it} + \gamma_i + \varepsilon_{it} \quad (2)$$

where i indexes the unit of observation and t indexes the year; left-hand side of the model refers to the six dependent variables mentioned earlier; $Treat_{it}$ is a dummy variable which equals 1 when the reviewer is in the treatment group (i.e., have been recruited to the Elite Squad) and 0 otherwise; $Status_{it}$ is a dummy variable which equals 1 if the observation occurs “after” the treatment start time and 0 otherwise; Cov_{it} is a vector of control variables; γ_i is the reviewer fixed effect; and ε_{it} is the error term. Based on the formula, we can estimate the short-term effect that we are interested in by examining the coefficient α_3 .

4.4.2. Long-term effect—To test the long-term effect, we used the matched three-year-long panel dataset and introduced a new variable *LongTerm*. Specifically, for those reviews written after being elite, we further split them into short-term reviews, i.e., those written in the first year after being elites, and long-term reviews, i.e., those written in the second year after being elites. Thus, for each treatment or control reviewer, her/his reviews were divided into three groups: before, after-short-term, and after-long-term. Under such setting, the DID model is:

$$DV_{ig} = \alpha_0 + \alpha_1 Treat_{ig} + \alpha_2 Status_{ig} + \alpha_3 Treat_{ig} \times Status_{ig} + \alpha_4 Treat_{ig} \times LongTerm_{ig} + \alpha_5 Cov_{ig} + \gamma_i + \varepsilon_{ig} \quad (3)$$

where i indexes a matched pair of reviewers, and g indexes the three groups. $Status_{ig}$ is a dummy variable which equals 1 if the observation corresponds to an “after” group (i.e., after-short-term, and after-long-term) and 0 otherwise; $LongTerm_{ig}$ is a dummy variable which equals 1 if the observation belongs to a “after-long-term” group and 0 otherwise. Given the DID model, the long-term effect is reflected by the coefficient α_4 .

5. Empirical Results

5.1. Main Results

Table 3 reports the estimation results of our short-term effect model (i.e., Equation 2) for the six different dependent variables. The time window considered here is one year before “being elite” and one year after “being elite”. With 721 reviewer pairs matched, we have 2,884 observations in total. As seen, all the coefficients of the key interactive term $Treat \times Status$ are significant and consistent with our hypotheses, suggesting behavioral changes observed after being incentivized with reevaluation mechanism. First, we can see that the “number of reviews” and “average review length” significantly increase after being elite, thus supporting H1a. This indicates that when a reviewer was recruited into the “Elite Squad” which is an extrinsic incentive with reevaluation mechanism, her/his contribution level increases due to increased accountability to the community. Specifically, the reviewer contributes about 34 more reviews in the first year after being elite and the average review length also increases by 21 more words. Regarding the opinions expressed, reviewers

become more conservative after having a high reputation and associated perks that are reevaluated every year. Particularly, they write reviews with higher average rating (increases by 0.168 stars), lower rating variance (decreases by 0.338), and lower extreme ratings (the ratio of one star decreases by 7.3%), suggesting H2a is supported. As for the readability, there is a significant decrease in Lexical Density, indicating higher readability. Hence, H3a is also supported.

Table 4 shows the estimation results of our long-term effect model (i.e., Equation 3) for the six dependent variables of interest. The time window we consider here is one year before “being elite”, one year after “being elite”, and two years after “being elite”. The long-term effects of incentives with reevaluation mechanism are reflected in the coefficients of the interactive term $Treat \times LongTerm$. First, results from the fixed effect DID model lend mixed support to H1b. When the dependent variable is *number of reviews*, the coefficient of the interactive term is negative and statistically significant, whereas the coefficient is positive and statistically significant when estimating the effect on *average review length*. Using *number of reviews* as an indicator of contribution level, elite reviewers lower their effort to write large amount of reviews in the second year of being elite due to the “complacency” effect after passing the first reevaluation process. However, the long-term effect on *average review length* is still positive but marginally decreasing compared to the first year (i.e., short term). The possible reason may be related to our definition of “long term”. Reviewers cannot write longer reviews enduringly and the average review length will keep at a stable level in long term. Limited by the time span of our collected data, we define “two years after being elite” as long term and observe positive effect of “being elite” on average review length. It would be an interesting future direction to explore how long this effect will last. Moreover, these two variables reflect the quantity and quality of contribution levels, respectively. We can also conclude that reviewers increase both the quantity and quality of reviews after being incentivized with reevaluation mechanism in short term, whereas focus more on the quality of reviews in long term.

Second, we find evidence in support of H2b. The coefficients of $Treat \times LongTerm$ are insignificant when dependent variables are *average rating*, *rating variance*, and *ratio of one star*. This shows that reviewers change their numerical ratings as a reaction of being recruited into “Elite Squad” only in short term. After they pass the first reevaluation process, their expressed opinions become stable. Third, H3b is not supported since the long-term effect on readability is still positive and statistically significant. The *average LD* for reviews decreases by 0.879, indicating a higher readability. Nevertheless, we can observe that the marginal effect on readability is decreasing, that is, the positive effect is stronger in the first year than that in the second year. Additionally, the third row of the table (i.e., $Treat \times Status$ as independent variable) also validates that our results from the long-term effect model as well as the short-term effect model (i.e., results in Table 3) are highly consistent.

5.2. Robustness Checks

Our main analyses thus far have consistently shown the effects of incentives with reevaluation mechanism on a series of reviewer behaviors such as contribution levels,

numerical ratings, and readability. To make the conclusions more convincing, we next conducted several robustness checks.

5.2.1. Alternative dataset—We used Tucson dataset to re-estimate our long-term effect model (i.e., Equation 3). It is a relatively smaller city compared to Phoenix and only has 572 reviewers in the “Elite Squad” community. After data filtering and propensity score matching, only 248 reviewer pairs are matched. Our results present consistent trends with our main findings. Specifically, we observed significant increases in contribution level, readability, and average ratings and decreases in rating variance and extreme ratings in short term. In addition, the incentive with reevaluation mechanism still has a long-term effect on contribution level and readability, yet no significant effect on numerical rating behaviors. Estimation results are shown in Table 5. This fact lends support to the robustness of our main findings.

5.2.2. Matching mechanism—To address self-selection issue as well as to identify the treated and control groups, we aligned the elite reviewers’ “treatment start time” and aggregated their review information in the last year before “treatment start time” for matching. One concern may arise from whether reviewers in the two groups have similar growth motivation or not. In other words, reviewers in treatment and control group may not be perfect matches if we consider their dynamic behaviors instead of behaviors only in the last year before “treatment start time”. To eliminate this concern, we implemented a different matching mechanism. The key idea is to match review history in the two years before the “treatment start time”, that is, variables used for matching include *average review length*($t - 2$), *review volume*($t - 2$), *average numerical rating*($t - 2$), *variance*($t - 2$), *ratio of extreme ratings*($t - 2$), *average LD*($t - 2$), *average number of votes*($t - 2$), *number of friends*($t - 2$), *average review length*($t - 1$), *review volume*($t - 1$), *average numerical rating*($t - 1$), *variance*($t - 1$), *ratio of extreme ratings*($t - 1$), *average LD*($t - 1$), *average number of votes*($t - 1$), and *number of friends*($t - 1$), where t is the treatment start time. The estimation results for Equation 3 are present in Table 6. Though being a better identification strategy, this alternative matching mechanism sacrifices in the number of data observations for model estimation. With that being said, we still observe qualitatively consistent results, as shown in the second and third row of Table 6.

5.2.3. Alternative outcome measures—As an extension, we explored alternative measure as outcome variable to further examine the treatment effect on extreme ratings. In our main analyses, the percentage of extreme ratings was operationalized as “ratio of one star”. As a robustness check, we used “ratio of five stars” to rerun the model and presented results in Table 7. Model 1 is estimated using Phoenix dataset, Model 2 is estimated using Tucson dataset, and Model 3 is estimated by matching two-year review history. We observed consistent results with the main findings, that is, reviewers become more conservative and lower the percentage of five stars in short term and stabilize in long term.

6. Discussion and Conclusions

Incentive mechanism has long been recognized as an important and effective means to induce users’ efforts on content sharing platforms, such as StackOverflow, Yelp, eBay, and

YouTube. Yet most of the existing research has focused on reputations that are everlasting (e.g., badges and virtual points) or financial rewards where no evaluation exists about the users' contributed content (e.g., rebates). There is still a significant gap in our understanding of how incentives with reevaluation mechanism actually influence users' behaviors including their contribution levels, the opinions they express, and how they express the opinions. To fill this gap, we draw from the accountability theory and propose three hypotheses to explore such effects in both short term and long term. Our research context is Yelp Elite Squad where reviewers with good reviewing history are awarded into the elite group and most importantly reevaluated each year. We designed a quasi-experimental setup by combining the propensity score matching (PSM) method with a difference-in-differences (DID) approach. To capture reviewers' behavioral changes, six dependent variables were developed to measure reviewing behaviors on three aspects, that is, contribution level, expressed opinions, and how to express the opinions. With these measures, we proposed and estimated a series of fixed effect DID models to examine how the incentives with reevaluation mechanism (i.e., "Elite") influence reviewers' behaviors. Our analyses find evidence that supports the proposed hypotheses. First, reviewers' contribution levels including the number of reviews and average review length increase significantly after "being elite" in short term, whereas have divergent trends in long term. Reviewers put more efforts on the reviews' quality (*average review length*) instead of quantity (*number of reviews*) in long term to maintain their benefits in the "Elite Squad". Second, elite members change the ratings of their reviews as a consequence of this affiliation only in short term. Particularly, they become more conservative and give less extreme reviews — they show higher average rating, lower rating variance, and lower percentage of extreme ratings. Third, the quality of reviews also significantly increases. Using readability as an indicator, we observe positive effect of the "Elite Squad" incentive on review readability in both short term and long term, and that the marginal effect in long term is decreasing.

The reported research is expected to make several contributions. First, our research extends the incentive literature by exploring the influence of reevaluation mechanism on reviewers' behaviors from three aspects: contribution level, numerical characteristics of reviews, and the quality of reviews. Second, we also propose a quasi-experimental design to further test whether this influence is a short-term or long-term effect. Third, we draw from the accountability theory to interpret reviewers' behaviors when they bear the pressure of maintaining the acquired benefits. It is clear that reviewers show different behavioral change patterns compared with those in contexts where no reevaluation exists, such as badge [15, 8, 14], incoming followers [19], or payments [32]. Concretely, despite that the traditional badge system is able to motivate users to contribute more before the badges are attained, user contribution levels drop significantly upon reaching the goals [14]. Our study, on the contrary, found that incentives with reevaluation mechanism can increase user contribution levels after users reach the goals (i.e., recruited to the Elite Squad). In context where users have more followers, they produce reviews in higher frequency and objectivity; however, their numeric ratings become more negative and more varied, which is different from our findings [19]. Last but not least, Qiao et al. [32] found that using monetary incentives greatly reduced the user's intrinsic motivations and led to long-term overjustification effect. Our research, however, demonstrates the long-term effectiveness of reevaluation mechanism in

motivating users to write high-quality reviews. It is worth noting that these studies are based on different contexts and hence the different behavioral change patterns may be attributed to factors more than the existence of reevaluation mechanism. More comprehensive empirical studies are expected to explore this interesting comparison in future research.

The present study also yields significant implications and actionable insights for business models that rely on user contributions, which finally enhances the decision making of relevant stakeholders. First, our reported study shows the effectiveness of incentives with reevaluation mechanism. This helps the UGC websites to design incentive mechanism with the hope of inducing more user efforts. Particularly, creating an elite dynamic community which needs to be reevaluated periodically is effective to increase user contribution levels and content quality. Though this positive effect is diminishing along with time, the reevaluation mechanism is still a good choice for designing incentive compared with the everlasting ones in which context users may drop their efforts significantly after achieving the goals. In light of our results on the linguistic features of reviews, such a reevaluation mechanism can induce more readable reviews in both short and long term, thereby providing more value to the UGC community in general. Thus, our findings provide important guidance for UGC platforms to make decisions about how to design incentives for the sake of the sustainability of platform. Second, voluntary contributions on product reviews sites serve as a reliable information source and significantly influence consumers' decision making. A prosperous community on Yelp stimulated by the reevaluation mechanism can better facilitate consumers' purchasing decisions. Third, because reviewers change their rating behaviors after being incentivized, relevant stakeholders need to be aware of the potential bias when making decisions based on the product reviews. In this case, although the incentive mechanism will facilitate the generation of more and higher-quality reviews, they might be biased. For UGC websites, they may need to balance the need to induce more user efforts and the need to avoid potential biases [19]. In addition to the incentive with reevaluation mechanism, the UGC websites can resort to some approaches from the literature (e.g., retailer-prompted reviews [55]) to mitigate the biases. For businesses that try to examine the market response to their products, or consumers that try to evaluate the products, they should carefully take into account such induced biases as well [19].

Our analysis is not without limitations and it can be extended in several directions. First, other indicators for the quality of reviews can be used to test H3, such as review helpfulness [56] and lexical richness [32]. Second, the elite reviewers can enjoy financial perks in addition to high reputation. They may be invited to private parties held in local businesses, which happens offline and is beyond our observation [57]. Given that the reevaluation mechanism is largely under-explored, this research focuses on the reevaluation mechanism regardless of the incentive types. In the future, we intend to explore this topic in finer granularity by separating the financial and non-financial incentives under the reevaluation mechanism. Third, the Yelp Elite Squad now classifies its members into three specialists, namely, writers, photographers, and adventurers. It is worth investigating the differences among these three types of elite reviewers and testing whether the modified incentive mechanism is more effective. Fourth, understanding reviewers' motivations of behavioral changes also has important implications for marketing practitioners but is well beyond the scope of our current data set. We may extend the study by conducting semi-structured

interviews or surveys to obtain respondent motivations in perceptual level. Finally, it would also be interesting and worthwhile to extend our framework to other contexts of UGC where users face the pressure of losing gained benefits and study how such a mechanism affects user behaviors.

Acknowledgments

This work was partly supported by the National Natural Science Foundation of China [71802024, 71621002], the Fundamental Research Funds for the Central Universities [2017QD009, YY19ZZB007], Chinese Academy of Sciences [ZDRW-XH-2017-3], and National Institutes of Health (NIH) of the USA [5R01DA037378-05].

Biography

Mingyue Zhang received her PhD degree in Management Science and Engineering from the School of Economic and Management, Tsinghua University, Beijing, China, in 2017. Her current research interests include incentive mechanism, recommender systems, and consumer behavior analysis. Her work has been published in journals such as *Decision Sciences*, *ACM TKDD*, *Decision Support Systems*, *Information Sciences*, etc.

Xuan Wei's research interests include crowd intelligence, false news detection, intention mining in social media, etc. He has published papers in journals and conference proceedings such as *ACM TKDD*, *International Journal of Intelligent Systems*, *PACIS*, and *ICDE*. He has presented his research at major conferences and workshops including *INFORMS*, *CIST*, and the *INFORMS Workshop on Data Science*. He received Best Paper Award in *WITS 2018*, and Best Paper Award Runner-up in *INFORMS Workshop on Data Science 2017*.

Daniel Dajun Zeng received the M.S. and Ph.D. degrees in industrial administration from Carnegie Mellon University and the B.S. degree in economics and operations research from the University of Science and Technology of China, Hefei, China. He is a Research Professor at the Institute of Automation in the Chinese Academy of Sciences and was formerly a Gentile Family Professor of MIS in the Department of Management Information Systems at the University of Arizona. His research interests include social computing, recommender systems, software agents, intelligence and security informatics, infectious disease informatics, and applied operations research and game theory. He has published one monograph as well as more than 300 peer-reviewed articles, and co-edited 22 books and proceedings. He served as the editor in chief of IEEE Intelligent Systems, the president of the IEEE Intelligent Transportation Systems Society, and the chair of INFORMS College on Artificial Intelligence. He also chaired many conferences including IEEE ISI, BioSecure, and SOCO. Now he serves as the editor in chief of ACM Transactions on Management Information Systems. He is a fellow of the IEEE and the AAAS.

References

- [1]. Luca M, Zervas G, Fake it till you make it: Reputation, competition, and yelp review fraud, *Management Science* 62 (12) (2016) 3412–3427.
- [2]. Hu N, Koh NS, Reddy SK, Ratings lead you to the product, reviews help you clinch it? The mediating role of online review sentiments on product sales, *Decision Support Systems* 57 (2014) 42–53.

- [3]. Berger J, Sorensen AT, Rasmussen SJ, Positive effects of negative publicity: When negative reviews increase sales, *Marketing Science* 29 (5) (2010) 815–827.
- [4]. Chintagunta PK, Gopinath S, Venkataraman S, The effects of online user reviews on movie box office performance: Accounting for sequential rollout and aggregation across local markets, *Marketing Science* 29 (5) (2010) 944–957.
- [5]. Luca M, Reviews, reputation, and revenue: The case of [Yelp.com](https://www.yelp.com), Available at SSRN: 10.2139/ssrn.1928601 (2016) 1–40.
- [6]. Sun M, How does the variance of product ratings matter?, *Management Science* 58 (4) (2012) 696–707.
- [7]. Zhu F, Zhang X, Impact of online consumer reviews on sales: The moderating role of product and consumer characteristics, *Journal of Marketing* 74 (2) (2010) 133–148.
- [8]. Li Z, Huang K-W, Cavusoglu H, Can we gamify voluntary contributions to online Q&A communities? Quantifying the impact of badges on user engagement, in: *Proc. of 2012 Workshop on Information Systems and Economics (WISE)*, 2012.
- [9]. Chen Y, Harper FM, Konstan J, Li SX, Social comparisons and contributions to online communities: A field experiment on MovieLens, *American Economic Review* 100 (4) (2010) 1358–1398.
- [10]. Pareto V, *Cours d'économie politique*, Vol. 1, Librairie Droz, 1964.
- [11]. Ba S, Whinston AB, Zhang H, Building trust in online auction markets through an economic incentive mechanism, *Decision Support Systems* 35 (2003) 273–286.
- [12]. Fayazi A, Lee K, Caverlee J, Squicciarini A, Uncovering crowdsourced manipulation of online reviews, in: *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM, 2015, pp. 233–242.
- [13]. Cabral L, Li LI, A dollar for your thoughts: Feedback-conditional rebates on eBay, *Management Science* 61 (9) (2015) 2052–2063.
- [14]. Goes PB, Guo C, Lin M, Do incentive hierarchies induce user effort? Evidence from an online knowledge exchange, *Information Systems Research* 27 (3) (2016) 497–516.
- [15]. Anderson A, Huttenlocher D, Kleinberg J, Leskovec J, Steering user behavior with badges, in: *Proceedings of the 22nd International Conference on World Wide Web*, 2013, pp. 95–106.
- [16]. Li L, Reputation, trust, and rebates: How online auction markets can improve their feedback mechanisms, *Journal of Economics & Management Strategy* 19 (2) (2010) 303–331.
- [17]. Li L, Xiao E, Money talks? An experimental study of rebate in reputation system design, *Management Science* 60 (8) (2014) 2054–2072.
- [18]. Kim C, Lin G, Bang H, Discovering Yelp elites: Reifying Yelp elite selection criterion, University of California-San Diego.
- [19]. Goes PB, Lin M, Au Yeung C.-m., “Popularity effect” in user-generated content: Evidence from online product reviews, *Information Systems Research* 25 (2) (2014) 222–238.
- [20]. Kahneman D, Tversky A, Prospect theory: An analysis of decision under risk, *Econometrica* 47 (2) (1979) 263–292.
- [21]. Milinski M, Semmann D, Krambeck H-J, Reputation helps solve the ‘tragedy of the commons’, *Nature* 415 (6870) (2002) 424. [PubMed: 11807552]
- [22]. Ryan RM, Deci EL, Intrinsic and extrinsic motivations: Classic definitions and new directions, *Contemporary Educational Psychology* 25 (1) (2000) 54–67. [PubMed: 10620381]
- [23]. Roberts JA, Hann I-H, Slaughter SA, Understanding the motivations, participation, and performance of open source software developers: A longitudinal study of the Apache projects, *Management Science* 52 (7) (2006) 984–999.
- [24]. Wasko MM, Faraj S, Why should I share? Examining social capital and knowledge contribution in electronic networks of practice, *MIS Quarterly* 29 (1) (2005) 35–57.
- [25]. Jabr W, Mookerjee R, Tan Y, Mookerjee VS, Leveraging philanthropic behavior for customer support: The case of user support forums, *MIS Quarterly* 38 (1) (2014) 187–208.
- [26]. Ye S, Gao G, Viswanathan S, Strategic behavior in online reputation systems: Evidence from revoking on eBay, *MIS Quarterly* 38 (4) (2014) 1033–1056.

- [27]. Fradkin A, Grewal E, Holtz D, The determinants of online review informativeness: Evidence from field experiments on Airbnb, Available at SSRN: <https://ssrn.com/abstract=2939064> (2017) 1–61.
- [28]. Cabral L, Hortacsu A, The dynamics of seller reputation: Evidence from eBay, *The Journal of Industrial Economics* 58 (1) (2010) 54–78.
- [29]. Resnick P, Zeckhauser R, Trust among strangers in Internet transactions: Empirical analysis of eBay's reputation system, in: *The Economics of the Internet and E-commerce*, Emerald Group Publishing Limited, 2002, pp. 127–157.
- [30]. Zhou M, Dresner M, Windle RJ, Online reputation systems: Design and strategic practices, *Decision Support Systems* 44 (2008) 785–797.
- [31]. Gneezy U, Meier S, Rey-Biel P, When and why incentives (don't) work to modify behavior, *Journal of Economic Perspectives* 25 (4) (2011) 191–210.
- [32]. Qiao D, Whinston AB, Lee S.-y., Incentive provision and pro-social behaviors, in: *50th Hawaii International Conference on System Sciences*, 2017, pp. 5599–5608.
- [33]. Lerner JS, Tetlock PE, Accounting for the effects of accountability, *Psychological Bulletin* 125 (2) (1999) 255. [PubMed: 10087938]
- [34]. Vance A, Lowry PB, Eggett DL, Increasing accountability through the user interface design artifacts: A new approach to addressing the problem of access-policy violations, *MIS Quarterly* 39 (2) (2015) 345–366.
- [35]. Bovens M, Two concepts of accountability: Accountability as a virtue and as a mechanism, in: *Accountability and European Governance*, Routledge, 2014, pp. 28–49.
- [36]. Vance A, Lowry PB, Eggett D, Using accountability to reduce access policy violations in information systems, *Journal of Management Information Systems* 29 (4) (2013) 263–290.
- [37]. Williams K, Harkins S, Latane B, Identifiability as a deterrent to social loafing: Two cheering experiments, *Journal of Personality and Social Psychology* 40 (2) (1981) 303–311.
- [38]. Hochwarter WA, Ferris GR, Gavin MB, Perrewé PL, Hall AT, Frink DD, Political skill as neutralizer of felt accountability—job tension effects on job performance ratings: A longitudinal investigation, *Organizational Behavior and Human Decision Processes* 102 (2) (2007) 226–239.
- [39]. Fandt PM, Ferris GR, The management of information and impressions: When employees behave opportunistically, *Organizational Behavior and Human Decision Processes* 45 (1) (1990) 140–158.
- [40]. Tetlock PE, Boettger R, Accountability: A social magnifier of the dilution effect, *Journal of Personality and Social Psychology* 57 (3) (1989) 388. [PubMed: 2778629]
- [41]. Staw BM, Knee-deep in the big muddy: A study of escalating commitment to a chosen course of action, *Organizational Behavior and Human Performance* 16 (1) (1976) 27–44.
- [42]. Schlenker BR, Weigold MF, Doherty K, Coping with accountability: Self-identification and evaluative reckonings., *Handbook of Social and Clinical Psychology: The Health Perspective*.
- [43]. Zhu L, Yin G, He W, Is this opinion leader's review useful? Peripheral cues for online review helpfulness, *Journal of Electronic Commerce Research* 15 (4) (2014) 267–280.
- [44]. Hennig-Thurau T, Gwinner KP, Walsh G, Gremler DD, Electronic word-of-mouth via consumer-opinion platforms: What motivates consumers to articulate themselves on the Internet?, *Journal of Interactive Marketing* 18 (1) (2004) 38–52.
- [45]. Baek H, Ahn J, Choi Y, Helpfulness of online consumer reviews: Readers' objectives and review cues, *International Journal of Electronic Commerce* 17 (2) (2012) 99–126.
- [46]. Cao Q, Duan W, Gan Q, Exploring determinants of voting for the “helpfulness” of online user reviews: A text mining approach, *Decision Support Systems* 50 (2) (2011) 511–521.
- [47]. Pan Y, Zhang JQ, Born unequal: A study of the helpfulness of user-generated product reviews, *Journal of Retailing* 87 (4) (2011) 598–612.
- [48]. Mudambi SM, Schuff D, What makes a helpful review? A study of customer reviews on [Amazon.com](https://www.amazon.com), *MIS Quarterly* 34 (1) (2010) 185–200.
- [49]. Centola D, The spread of behavior in an online social network experiment, *Science* 329 (5996) (2010) 1194–1197. [PubMed: 20813952]

- [50]. Keegan J, Kabanoff B, Indirect industry- and subindustry-level managerial discretion measurement, *Organizational Research Methods* 11 (4) (2008) 682–694.
- [51]. Read J, *Assessing vocabulary*, Cambridge University Press, 2000.
- [52]. Becker S, Ichino A, Estimation of average treatment effects based on propensity scores, *Stata Journal* 2 (4) (2002) 358–377.
- [53]. Rosenbaum PR, Rubin DB, The central role of the propensity score in observational studies for causal effects, *Biometrika* 70 (1) (1983) 41–55.
- [54]. Hosanagar K, Fleder D, Lee D, Buja A, Will the global village fracture into tribes? Recommender systems and their effects on consumer fragmentation, *Management Science* 60 (4) (2013) 805–823.
- [55]. Askalidis G, Kim SJ, Malthouse EC, Understanding and overcoming biases in online review systems, *Decision Support Systems* 97 (2017) 23–30.
- [56]. Korfiatis N, García-Bariocanal E, Sánchez-Alonso S, Evaluating content quality and helpfulness of online product reviews: The interplay of review helpfulness vs. review content, *Electronic Commerce Research and Applications* 11 (3) (2012) 205–217.
- [57]. Askay DA, Gossett L, Concealing communities within the crowd: Hiding organizational identities and brokering member identifications of the Yelp elite squad, *Management Communication Quarterly* 29 (4) (2015) 616–641.

Highlights

- We draw from the accountability theory to investigate the influence of incentives with reevaluation mechanism on reviewers' behavior in content sharing platforms.
- We use propensity score matching and difference-in-differences method to analyze the data collected from Yelp platform.
- In short term, reviewers significantly increase their contribution levels after being incentivized with reevaluation mechanism, become more conservative, and also increase the readability of their reviews.
- In long term, reviewers continue to improve the quality of reviews while their numerical rating behaviors stabilize.

Figure 1:
Distribution of the Number of Reviewers

(a) Num. of Reviews

(b) Avg. Review Length

(c) Avg. Rating

(d) Variance of Ratings

(e) Ratio of One-star Ratings

(f) Avg. LD

Figure 2:
Elite Reviewers' Behaviors Across Four Years

Figure 3:
Distribution of Propensity Score Before and After Matching (for short-term effect model)

Table 1:

Descriptive Statistics of the Dataset

	Phoenix (119, 277 reviewers, 546, 505 reviews in total)						
	Registration time	Elite duration	#friends	#followers	Numerical ratings	Review length	Review LD
Mean	NA	0.023	74.07	0.16	3.90	108.73	77.85
Std. Dev	NA	0.32	109.82	1.39	1.38	104.40	12.04
Minimum	Apr. 2006	0	1	0	1	0	0
Maximum	Sep. 2018	10 years	3, 687	168	5	1, 035	100
	Tucson (53, 964 reviewers, 182, 064 reviews in total)						
Mean	NA	0.025	65.65	0.104	3.88	106.72	77.95
Std. Dev	NA	0.297	93.71	0.949	1.37	100.13	11.88
Minimum	Apr. 2005	0	1	0	1	0	0
Maximum	Sep. 2018	10 years	1, 397	160	5	1, 003	100

Table 2:

Statistical Tests Before and After Matching (for short-term effect model)

		Before Matching		After Matching	
	Elite	Non-elite	<i>p</i> – value	Non-elite	<i>p</i> – value
Obs.	721	18213	NA	721	NA
NbrReviews	32.05 (48.39)	10.51 (15.15)	< 0.001	30.83 (50.27)	0.640
AvgNbrVotes	3.04 (4.41)	2.08 (3.17)	< 0.001	2.98 (6.82)	0.849
AvgReviewLength	119.97 (64.87)	98.07 (67.40)	< 0.001	126.15 (80.10)	0.108
AvgRating	3.96 (0.49)	3.85 (0.81)	< 0.001	3.97 (0.47)	0.543
Variance	1.19 (0.78)	1.48 (1.15)	< 0.001	1.13 (0.86)	0.173
RatioOneStar	0.07 (0.09)	0.14 (0.19)	< 0.001	0.06 (0.10)	0.153
AvgLD	76.23 (7.15)	79.28 (7.99)	< 0.001	75.27 (8.30)	0.180
NbrFriends	180.80 (260.56)	79.92 (133.95)	< 0.001	172.62 (271.59)	0.560

Notes: Standard deviations are in parentheses.

Table 3:

Results of Short-Term Effects

Variables	Num. of reviews	Avg. review length	Avg. rating	Rating variance	Ratio of one star	Readability: LD
<i>Treat</i> × <i>Status</i>	34.280 *** (2.887)	21.150 *** (3.118)	0.168 *** (0.030)	−0.338 *** (0.049)	−0.073 *** (0.007)	−3.112 *** (0.342)
<i>Treat</i>	86.240 (119.8)	−16.370 (129.4)	1.072 (1.225)	−0.755 (2.018)	0.150 (0.290)	−4.727 (14.190)
<i>Status</i>	−7.311 ** (2.414)	−2.034 (2.607)	−0.149 *** (0.025)	0.210 *** (0.041)	0.045 *** (0.006)	0.664 * (0.286)
Fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Num. of obs.	2884	2884	2884	2884	2884	2884
Adjusted R^2	0.3447	0.6598	0.4479	0.3974	0.3062	0.6315

Notes:

 $p < 0.001$ **
 $p < 0.01$ *
 $p < 0.05$ +
 $p < 0.1$.

Robust standard errors are shown in parentheses.

Table 4:

Results of Long-Term Effects

Variables	Num. of reviews	Avg. review length	Avg. rating	Rating variance	Ratio of one star	Readability: LD
<i>Treat</i> × <i>LongTerm</i>	−15.76 ^{***} (3.249)	8.929 [*] (3.982)	0.049 (0.036)	−0.036 (0.061)	−0.003 (0.008)	−0.879 [*] (0.406)
<i>Treat</i> × <i>Status</i>	37.330 ^{***} (3.225)	20.88 ^{***} (3.953)	0.106 ^{**} (0.036)	−0.226 ^{***} (0.060)	−0.060 ^{***} (0.008)	−2.921 ^{***} (0.403)
<i>Treat</i>	12.120 (48.95)	−271.3 ^{***} (60.00)	0.219 (0.545)	−0.418 (0.912)	−0.144 (0.128)	23.872 ^{***} (6.124)
<i>Status</i>	−9.483 ^{***} (2.537)	−0.039 (3.110)	−0.101 ^{***} (0.028)	0.092 ⁺ (0.047)	0.033 ^{***} (0.006)	0.419 (0.317)
Fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Num. of obs.	2895	2895	2895	2895	2895	2895
Adjusted R^2	0.4363	0.6737	0.4866	0.3728	0.3182	0.6598

Notes:

^{***}
 $p < 0.001$ ^{**}
 $p < 0.01$ ^{*}
 $p < 0.05$ ⁺
 $p < 0.1$.

Robust standard errors are shown in parentheses.

Table 5:Robustness I: Using *Tucson* Dataset

Variables	Num. of reviews	Avg. review length	Avg. rating	Rating variance	Ratio of one star	Readability: LD
<i>Treat</i> × <i>LongTerm</i>	−7.414 [*] (3.175)	11.760 [*] (5.489)	0.028 (0.051)	0.008 (0.082)	−0.008 (0.012)	−1.209 [*] (0.576)
<i>Treat</i> × <i>Status</i>	26.100 ^{***} (3.154)	19.780 ^{***} (5.452)	0.080 ⁺ (0.050)	−0.247 ^{**} (0.082)	−0.047 ^{***} (0.012)	−3.047 ^{***} (0.573)
Fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Num. of obs.	1488	1488	1488	1488	1488	1488
Adjusted R^2	0.3909	0.6703	0.4634	0.3709	0.3277	0.6468

Notes:

 $p < 0.001$ **
 $p < 0.01$ *
 $p < 0.05$ ⁺
 $p < 0.1$.

Robust standard errors are shown in parentheses.

Table 6:

Robustness II: Alternative Matching Mechanism

Variables	Num. of reviews	Avg. review length	Avg. rating	Rating variance	Ratio of one star	Readability: LD
$Treat \times LongTerm$	-10.13^{**} (3.699)	10.108^* (5.171)	-0.010 (0.054)	0.049 (0.088)	0.005 (0.012)	-1.354^* (0.555)
$Treat \times Status$	26.32^{***} (3.666)	13.250^{**} (5.125)	0.155^{**} (0.053)	-0.256^{**} (0.087)	-0.064^{***} (0.012)	-1.870^{***} (0.550)
Fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Num. of obs.	1332	1332	1332	1332	1332	1332
Adjusted R^2	0.5876	0.7218	0.4583	0.3921	0.3824	0.6978

Notes:

 $p < 0.001$ **
 $p < 0.01$ *
 $p < 0.05$ +
 $p < 0.1$.

Robust standard errors are shown in parentheses.

Table 7:

Robustness III: “Ratio of Five Stars” as DV

Variables	Model 1	Model 2	Model 3
<i>Treat</i> × <i>LongTerm</i>	0.016(0.0149)	0.007(0.021)	0.005(0.023)
<i>Treat</i> × <i>Status</i>	−0.052 ^{***} (0.014)	−0.067 ^{**} (0.021)	−0.044 [*] (0.022)
Fixed effects	Yes	Yes	Yes
Num. of obs.	2895	1488	1332
Adjusted R^2	0.5933	0.5274	0.5440

Notes:

^{***}
 $p < 0.001$
^{**}
 $p < 0.01$
^{*}
 $p < 0.05$
⁺
 $p < 0.1$.

Robust standard errors are shown in parentheses.