The Effectiveness of Non-Persistent Social Status as an Incentive Mechanism

Completed Research Paper

Xuan Wei **Mingyue Zhang**

Daniel Dajun Zeng

Abstract

Content sharing platforms such as product review websites largely depend on users' voluntary contributions. In order to motivate users to contribute more, many platforms established reputation-based incentive mechanisms. Though the academic community has devoted much effort to study the effectiveness of these mechanisms, most of the existing research has focused on everlasting reputations such as badges and points. It's still largely underexplored how non-persistent social status actually influences user's behavior ranging from contribution level, the opinion they express, to how they express. In this paper, we answer this question by examining data from Yelp Elite Squad where reviewers with good reviewing history are awarded into the elite group and most importantly re-evaluated each year. By applying propensity score matching and difference-in-difference methods, we find that in short term, reviewers significantly increase their contribution levels after acquiring the non-persistent social status, become more conservative with lower percentage of extreme ratings, and also increase the readability of their reviews. In long term, they continue to improve the quality of reviews while their numerical rating behaviors stabilize. Our research has significant implications and actionable insights for business models that rely on user contributions.

Keywords: Incentive mechanism, product reviews, non-persistent social status, reputation maintenance, accountability theory

Introduction

An increasing number of websites today heavily rely on user-generated content (UGC). The voluntary contributions from the community are the core power of retaining existing users, as well as attracting new ones (Li et al. 2012). These content sharing platforms range from e-commerce sites, social network sites, blogs, video or image sharing sites, and many others, from which the product reviews sites are one of the most common platforms. For example, Yelp, TripAdvisor, and Angie's List have become increasingly popular in the past few decades (Luca and Zervas 2016), all of which are pure non-commercial review sites. There is also mounting evidence indicating that online reviews significantly influence consumer choices and product sales (Berger et al. 2010; Chintagunta et al. 2010; Luca 2016; Luca and Zervas 2016; Sun 2012; Zhu and Zhang 2010). 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?

In order to motivate users to contribute more, many platforms have established reputation-based incentive mechanisms in various forms such as badges, points, or qualification (Anderson et al. 2013; Goes et al. 2016). The proliferation of these incentive mechanisms has drawn much attention from the academic community where they examine whether the incentive mechanisms are effective in inducing users to contribute more and higher quality contents. Most of the relevant research has focused on reputations that are *everlasting*, such as badges and points (Anderson et al. 2013; Goes et al. 2016). In that context, users can always keep the glory badges or higher ranks once acquiring them, even though they may stop contributing. However, it's quite common that the reputations are not everlasting. Take Yelp as an example, the Yelp Elite Squad program was launched in 2005, and serves as one way of recognizing people who are active in the Yelp community and role models on and off the site. The program is a yearly program and users who were admitted into the Elite Squad need to be re-evaluated every year. Another example is Amazon Turk, where the site automatically grants the Masters qualification to users based on statistical models that analyze workers' historical performance. Notably, the Masters qualification can be revoked if the worker's performance drops and he or she no longer scores highest across requester-provided and marketplace data points[1].

Unfortunately, there is little research exploring how such non-persistent social status actually affects user behaviors. This non-persistent social status incentive is very different from the traditional badges or points accumulating programs, despite some apparent similarities. In traditional settings, users bear no pressure of losing the glory, while users need to continue contributing more and high quality contents in order to maintain their status in the non-persistent situation. Given the wide spread, importance, and uniqueness of the non-persistent scenario, it is urgent that we systematically examine whether and how the incentive with non-persistent social status influences user behaviors.

In this paper, we attempt to fill the observed research gap by studying how reviewers in the Yelp Elite Squad change their behaviors after they get the high social status (i.e., being elites). This issue is challenging because there are two potential counterarguments. On one hand, such incentives could be treated as a goal for reviewers and their contribution levels may drop significantly after reaching the goal (i.e., being elites) due to "complacency" effects (Goes et al. 2014, 2016). On the other hand, the Elite status is prominently displayed next to elite users' names on the users' profile avatars, as well as all reviews that they wrote. Thus, it signifies the user 's status in the community and distinguishes them from their peers. Drawing from prospect theory (Kahneman and Tversky 1979), users 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, which in some sense, is consistent with previous literature indicating that reputation and recognition motivate people to behave in a better way (Goes et al. 2014; Milinski et al. 2002). Also, if there is any change, it's still not clear whether the change is temporary or long lasting. Hence, we address the following research question:

How does the incentive mechanism with non-persistent social status influences user 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?

In the reported study, we first develop our hypotheses based on the accountability theory. We use Yelp Elite Squad as an example to test the proposed hypotheses. Specifically, we collected the profiles of elite reviewers in Yelp and analyzed their behavior 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 also collected the profiles of non-elite reviewers, and combined 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 non-persistent social status on the reviewer's behavior.

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

[1] https://www.mturk.com/worker/help#what_is_master_worker

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.

Related Literature

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 tend to build various incentive mechanisms to increase the users' extrinsic motivations (Ryan and Deci 2000). It might be financial rewards such as free gifts (Fayazi et al. 2015), rebates (Cabral and Li 2015), and extra payments (Roberts et al. 2006), or nonfinancial rewards such as higher social status (Goes et al. 2016; Roberts et al. 2006; Wasko and Faraj 2005), social comparison (Chen et al. 2010; Jabr et al. 2014), and achievement badges (Anderson et al. 2013; Li et al. 2012). There is also evidence showing that users' contribution behavior is strongly influenced by these incentive mechanism (Gneezy et al. 2011; Jabr et al. 2014).

Researchers have studied this issue in different contexts (i.e. UGC sites). (1) In open source software communities, Roberts et al. (2006) studied the effects of both financial and nonfinancial 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) 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) 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) identified a short-term positive effect of winning new badges in Q&A communities. (3) In product review sites and e-commerce sites where review is an important component, Goes et al. (2014) examined the change of reviewer behaviors under the website's mechanism that allows one user to subscribe to another. They found that the subscription mechanism was effective in inducing user efforts, that is, users produce more reviews and more objective reviews as they become more popular. Qiao et al. (2017) argued that monetary incentive provision in review platforms would greatly damage the users' original altruistic and intrinsic motivations and result in lower quality and less helpful reviews in short term.

The reputation-based incentives in above studies are similar to the one we study since they all highlight users' status. However, users bear no pressure of losing glory, either badges or points. Similarly, reviewers in epinions.com still bear little pressure of losing popularity. The Yelp Elite Squad in our study, by contrast, is a non-persistent social status for reviewers where reviewers need to be re-evaluated every year. It is expected that reviewers behave differently under the non-persistent reputation mechanism. Hence, findings in existing studies do not necessarily apply, and our research further contributes to the growing literature.

Accountability theory

The accountability theory was originally developed by Lerner and Tetlock (1999), and then widely applied in a variety of fields, including psychology, philosophy, ethics, and organizational behavior (Vance et al. 2015). It is a process in which a person has a potential obligation to explain her/his activities toward another party who can make judgement on these activities and also to administer potential positive or negative consequences as a response to them (Bovens 2010). Thus, the two key elements to stimulate accountability perceptions are overt expectations of evaluation and awareness of monitoring. Later, Vance et al. (2013, 2015) extended the accountability theory to IS context and developed four user-interface design artifacts (i.e., identifiability, expectation of evaluation, awareness of monitoring, and social presence) to raise users' accountability perceptions within systems. Identifiability refers to a person's "knowledge that her/his outputs could be linked to her/him" and thus reveals her/his true identity (Williams et al. 1981). Therefore, individuals who perceive increased identifiability know that they can be made responsible for their actions (Lerner and Tetlock 1999). Second, expectation of evaluation is the belief that one's activities will be assessed and judged by others with some implied consequences (Lerner and Tetlock 1999). Such awareness will drive socially

desirable behaviors (Hochwarter et al. 2007). 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. Therefore, 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 prosocial behaviors (Fandt and Ferris 1990), increase conformity to expected behaviors (Tetlock and Boettger 1989), increase conservatism (Staw 1976), and decrease risk taking (Schlenker et al. 1991).

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 expressed.

Status among peers is a powerful motivator (Anderson et al. 2013; Goes et al. 2016), resulting in higher respect and admiration from peers. This is manifested in our scenario from two aspects. First, entering the "Elite Squad" reflects high social status of the reviewers because of the tiny minority nature of this group (Askay and Gossett 2015). Studies show that individuals who are awarded higher status by their teammates respond by contributing even more to the group than they had in the past (Ariely et al. 2009; Willer 2009). Second, there are certain restrictions for members in Elite Squad. For instance, users are required to disclose their real names and post real photos in the profile from which they experience more identifiability and individuation (Vance et al. 2015). According to the accountability theory, these effective cues to identity create a sense of accountability (Askay and Gossett 2015; Vance et al. 2015), resulting in increased pro-social behaviors (Fandt and Ferris 1990). Meanwhile, Elite Squad memberships are renewed each calendar year. Thus, users have a clear expectation of evaluation which is another important component of accountability.

Serving as the role of opinion leaders, members of Elite Squad are also more likely to be recognized than the average users in the virtual community, as well as occupying a structurally advantageous position within the social network (Zhu et al. 2014). As a result, users have more exposure chances to the public after they get the high social status (i.e., being elite). It is natural to expect that users with a larger online audience (i.e., social presence) should be more likely contribute more reviews(Goes et al. 2014; Jabr et al. 2014). Notably, the findings that Goes et al. (2016) reached building upon goal-setting theory is not suitable here, that is, users exert more effort before reaching goals and reduce efforts significantly once they reach goals. The Elite Squad program implements a nomination mechanism, that is, the reviewers in this Squad are selected by a "Skull and Bones-like process" and is "proffered by a governing body known as The Council". Thus, there is no pre-defined, specific goal for reviewers and they don't know how much distance they are from the elite status. For the long term effect, the encouragement effect of high status and accountability should be stronger at the beginning when they were first admitted into this Squad (Goes et al. 2014). Using *review length* as the indicator of contribution level, the increasing rate should decrease until it reaches a stable state. For the *number of reviews*, there may be certain "complacency" effect after the reviewers passing the 1st re-evaluate process (i.e., the second year after being elite), thus decreasing in long term. We therefore hypothesize the following:

> *H1: (a) After receiving the non-persistent social status, reviewers will increase their contribution levels such as the number of reviews and length of reviews in short term, compared with those not receiving high status. (b) In long term, the number of reviews will decrease while the length remains stable.*

The opinions that user expresses may also be influenced by the changing of social status, which can be reflected by the numerical ratings. Here we focus on three aspects of ratings: the average rating, variance and ratio of extreme ratings. As mentioned earlier, elite users experience high identifiability and expectation of evaluation, as well as a high social presence compared with those non-elite users. As a consequence of this affiliation, users exhibit discomfort with these public expressions of identification

and also change the content and ratings of their reviews (Askay and Gossett 2015). Specifically, when individuals are performing identifiable behaviors, they are more likely to engage in a systematic processing instead of heuristic processing, eliciting an increased sense of accountability (Vance et al. 2015). The similar process is aroused when users are aware of being re-evaluated every year and have a high perception of social presence in the community. With accountability towards the system, users will increase conservatism (Staw 1976) and decrease risk taking (Schlenker et al. 1991). Thus, elite reviewers will give fewer negative ratings to mitigate risks, resulting in higher average rating. Meanwhile, the rating extremity is a large degree of deviation from the average rating of all reviews (Baek et al. 2012; Cao et al. 2011; Pan and Zhang 2011; Zhu et al. 2014). Previous research found that moderate messages could enhance source credibility (Mudambi and Schuff 2010) while rating extremity diluted the influence of reviewer credibility (Zhu et al. 2014). Thus, elite users will write fewer reviews with extreme ratings. Since all online product reviews follow an asymmetric bimodal distribution with J shape, the five-star ratings are not normally considered extreme ratings. Hereby, we focus on the ratio of extreme negative ratings (i.e., one star) in main analysis. Reviewers with high status tend to write fewer extreme reviews to increase their credibility. As a consequence, the rating variance will also decrease after being elite. In long term, elite reviewers still possess high accountability due to the existence of identity and re-evaluation requirements, which differentiates with the goal-driven process in traditional badges mechanism. Thus, reviewers will keep all the numerical rating behaviors to mitigate the risk. The marginal effects on numerical ratings are decreasing until stabilized. We put forward the hypothesis:

> *H2: (a) After receiving the non-persistent social status, reviewers will write reviews with significantly different opinions in short term, including higher average ratings, lower rating variance, and lower percentage of extreme ratings. (d) In long term, these numerical rating behaviors remain stable.*

To measure the quality of reviews, a natural metric is the readability of review contents (Goes et al. 2014). Reviewers who receive recognition of Elite may be given a label that signals a type of connoisseurship or expertise, which could fulfill one's self-enhancement need (Hennig-Thurau et al. 2004). As experts in Yelp community, the elite users tend to review in higher quality and frequency to maintain the reputation, which in some sense, is consistent with previous literature indicating that reputation and recognition help people behave in a better way (Goes et al. 2014; Milinski et al. 2002). The increased perception of accountability also elicit reviewers' conformity to expected behaviors (Tetlock and Boettger 1989), thus writing reviews with high quality. Similar to the numerical rating behaviors, the marginal effect on readability will decrease until stabilized. Thus, we hypothesize:

> *H3: (a) After receiving the non-persistent social status, reviewers will write reviews with higher readability in short term. (b) In long term, the readability remains stable.*

Research Design

Data Collection

We collected the reviewers' information from Yelp[2] on October 2018. Since it is not feasible to collect all the data across the website, we focus on reviewers from certain cities, namely, Phoenix and Tucson in United States. In order to collect the information about reviewers, we use a snowball crawling strategy. For each city, we started from the Community Manager[3] and collected his/her first-degree friends who were also in the same city. The similar process was repeated for the collected friends until we got the six-degree friends. Figure 1 presents the number of elite and non-elite reviewers as a function of degree in Phoenix and Tucson, respectively. We can observe that the elite reviewers are concentrated upon the first and second degree to the community manager and there are no more new reviewers in the

[2] https://www.yelp.com

[3] Community managers are on-the-ground Yelp ambassadors who curate the Local Yelp newsletter, host incredible events for the active Yelp communities.

network after six-degree collection. Hence, we claim that we have collected almost all reviewers including both elites and non-elites in Phoenix and Tucson[4].

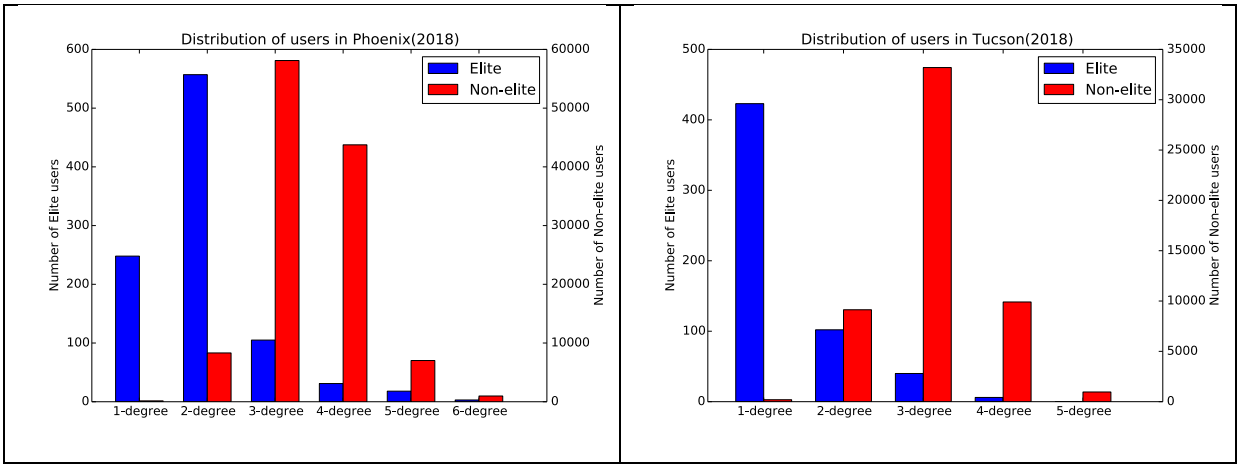

Figure 1. Distribution of the Number of Reviewers

In total, 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 users are just the minority group in Yelp such that only less than 2% of the reviewers are in Elite Squad in our collected dataset[5]. This is another significant difference compared with other incentive mechanism such as badges, ranks, and points.

For each reviewer, we collected the year s/he was elite (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 in the *n*-th degree of the friend network of the Community Manager. We kept tracking such information because exposure to more elite reviewers may change reviewer's behavior due to the peer influence (Centola 2010), 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 users, and 390,510 written by non-elite users. For the Tucson dataset, we got 182,064 reviews in total with 67,276 written by elite users and 114,788 written by non-elite users. 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) (Keegan and Kabanoff 2007). The calculation formula is as follows[6].

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

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.

Matching

One concern of examining the behavioral change of entering Elite Squad is that it may be attributed to inherent user characteristics rather than acquiring higher social status. That is to say, users 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 (Becker and Ichino 2002; Rosenbaum and Rubin 1983) method. To this end, we need to construct a "control group" with reviewers who never acquire the Elite status 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 match. First, reverse causality and simultaneity bias are mitigated in the panel data because we match the reviewers

[4] For those isolated users that are not connected to anyone, excluding them has no influence on our analyses.

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

[6] The measurement is negatively correlated with *readability*.

Table 1. Descriptive Statistics of the Dataset

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

by merely using activities before the "treatment start time". Second, since the Elite program adopts a nomination mechanism, it is possible to find a control group because not all users enter Elite Squad automatically after achieving certain contribution levels or having certain behaviors.

To test the short-term effect, we apply the matching method in the following manner. As mentioned above, one challenge for matching is that elite-reviewers do not necessarily enter the elite program at the same year. We need to match treated users (Elite users) with control users (non-Elite users) prior to the "treatment start time" (being elite) for each user, but we do not have a well-defined "treatment start time" for the control users. For treatment group, we first filter elite users to keep those having at least one year reviewing history after being elite and then align the year that they became a member of "Elite Squad". We then aggregate reviewing information of each reviewer for one year before being elite (t=0) and the year after being elite (t=1). When conduct matching, we only use data prior to the "treatment start time" (i.e., t=0) for each user in treatment group. For control group, those reviewers that have never been elite are potential matches. We first filter the non-elites with less than two years reviewing history. This is because we need to use the first year for matching and the second year for comparison to test the short-term effect in following steps. By leaving the last year out for potential matches and aggregating the review information for remaining each year, we perform a Nearest Neighboring matching by 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 the control group user and her/his following year's reviewing information is then reserved for following difference-in-difference estimation. Finally, a two-year-long panel dataset for both the treatment group and control group before and after treatment time is constructed and 721 pairs have been matched.

To ensure that our matching is successful, we plot the distribution of propensity score before and after matching between two groups in Figure 2. 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. We further conduct statistical tests and conclude that the distributions of all variables are not significantly different between control and treatment group after each matching. Due to space constraints, we omit the details.

The matching procedures for testing the long-term effect are quite similar 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 conduct matching, we only use the aggregate information of the year before being elite. For control group, we filter the non-elites 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 examination of the short-term and long-term effects, respectively. As a result, 485 pairs are matched, and we obtain a three-year-long panel dataset for both the treatment and control group before and after treatment time. Statistical tests are also conducted to

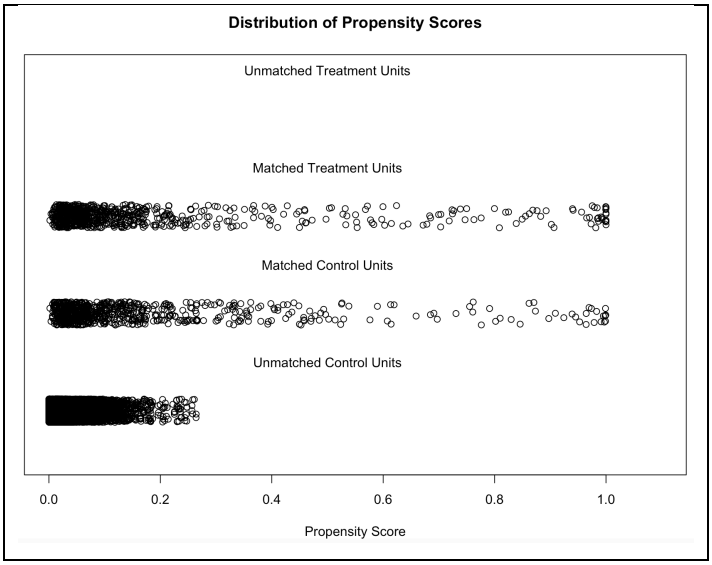

Figure 2. Distribution of Propensity Score Before and After Matching

show there is no significant differences between variables after matching. Noted that the reason for performing two matchings is to retain data observations as many as possible.

Difference-in-Difference Estimation

Short-term effect

Once matched pairs are identified, we extract their activities *before* ($Status = 0$) and *after* ($Status = 1$) the "being elite" event occurs and then compare them in a difference-in-difference manner to account for factors such as time trends and maturation that might be confounded with elite status in a one-group pre–post design (Hosanagar et al. 2014). 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 develop 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 choose 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 use 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 specify our DID model as follows:

$$DV_{it} = \alpha_0 + \alpha_1 Treat_{it} + \alpha_2 Status_{it} + \alpha_3 Treat_{it} \times Status_{it} + \alpha_4 Cov_{it} + \gamma_i + \varepsilon_{it} \qquad (1)$$

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. After fitting the model, we can estimate the short-term effect that we are interested in by examining the coefficient α_3.

Long-term effect

To test the long-term effect, we use the matched three-year-long panel dataset and introduce a new variable $LongTerm$. Specifically, for those reviews written *after* being elite, we further split them into *short-term reviews* (i.e., reviews in the first year after being elites) and *long-term reviews* (i.e., reviews in the second year after being elites). We can call these 3 splits different *groups*. Thus, for each treatment or control reviewer, we have three groups of reviews: before, after-short-term, and after-long-term. Therefore, the DID model is:

$$DV_{ig} = \alpha_0 + \alpha_1 Treat_{ig} + \alpha_2 Status_{ig} + \alpha_3 Treat_{ig} \times Status_{ig} + \alpha_4 Treat_{ig} \times LongTerm_{ig} + \alpha_5 Cov_{ig} + \gamma_i + \varepsilon_{ig} \quad (2)$$

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 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 $(\alpha_3 + \alpha_4)$. Noted that since the control group didn't acquire Elite status, we assume there is no significant difference between $DV(control\_after\_shortTerm)$ and $DV(control\_after\_longTerm)$. Particularly,

$$\alpha_3 = [DV(treatment\_after\_shortTerm) - DV(control\_after\_shortTerm)] - [DV(treatment\_before) - DV(control\_before)]$$

$$\alpha_3 + \alpha_4 = [DV(treatment\_after\_longTerm) - DV(control\_after\_shortTerm)] - [DV(treatment\_before) - DV(control\_before)]$$

Results

Table 2 reports the estimation results of our short-term effect model (i.e., Equation (1)) for the six different dependent variables. The time window we consider 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 acquiring high social status. 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", 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 expands 21 more words. Regarding to the opinions expressed, reviewers become more conservative after having a high social status and aware of being re-evaluated 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 0.073%), 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 2. Results of Short-Term Effects

Variables	Num. of reviews	Avg. review length	Avg. rating	Rating Variance	Ratio of one star	Readability: LD
$Treat \times Status$	34.280*** (2.887)	21.150*** (3.118)	0.168*** (0.030)	-0.338*** (0.049)	-0.073*** (0.007)	-3.112*** (0.342)
$Treat$	86.240 (119.8)	-16.370 (129.4)	1.072 (1.225)	-0.755 (2.018)	0.150 (0.290)	-4.727 (14.190)
$Status$	-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 3 shows the estimation results of our long-term effect model (i.e., Equation (2)) for the six outcomes of interest respectively. 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 non-persistent social status 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: the coefficient of the interactive term in

terms of "*number of reviews*" is negative and statistically significant, whereas the coefficient of estimating the effect on "*average review le*ngth" is positive and statistically significant. 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 re-evaluate 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. Due to the limitation of time span, 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 acquiring the non-persistent social status 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 re-evaluate process, their expressed opinions become stabilized. Third, H3b is not supported since the long-term effect on readability is still positive and statistically significant. The average Lexical Density for reviews decreases by 0.879, indicating a higher readability. Nevertheless, we can also 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 also validates that our results from the long-term effect model as well as the short-term effect model are highly consistent.

Table 3. Results of Long-Term Effects

Variables	Num. of reviews	Avg. review length	Avg. rating	Rating Variance	Ratio of one star	Readability: LD
$Treat \times LongTerm$	-15.76*** (3.249)	8.929+ (3.982)	0.049 (0.036)	-0.036 (0.061)	-0.003 (0.008)	-0.879* (0.406)
$Treat \times Status$	37.330*** (3.225)	20.88*** (3.953)	0.106** (0.036)	-0.226*** (0.060)	-0.060*** (0.008)	-2.921*** (0.403)
$Treat$	12.120 (48.95)	-271.3*** (60.00)	0.219 (0.545)	-0.418 (0.912)	-0.144 (0.128)	23.872*** (6.124)
$Status$	-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.

Our main analyses thus far have consistently shown the effects of non-persistent social status on a series of reviewer behaviors, such as contribution level, numerical ratings, and readability. We also conducted several robustness checks to validate the results, including using alternative datasets (i.e., Tucson), implementing a different matching mechanism, and using different dependent variables measurement. We observed consistent results with the main analyses. The details are omitted due to space constraints.

Discussion and Conclusions

Reputation-based incentive mechanism has long been recognized as an important and effective means to induce users' efforts on content sharing platforms. Yet most of the existing research has focused on reputations that are everlasting, such as badges and points. There is still a significant gap in our understanding of how non-persistent social status actually influences user behaviors such as their

contribution level, the opinions they express, and how they express. 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 built on Yelp Elite Squad where users with good reviewing history are awarded into the elite sub-community and most importantly re-evaluated each year. We design a quasi-experimental setup by combining the propensity score matching (PSM) method with a difference-in-difference (DID) approach. Our analyses find evidence that supports the proposed hypotheses. First, users' 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 identity 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 such that they show higher average rating, lower rating variance and lower percentage of extreme ratings. Third, the quality of reviews also has a significant increase. Using readability as an indicator, we observe positive effect of the "Elite" status on review readability in both short term and long term, and the marginal effect in long term is decreasing.

The research is expected to make several contributions. First, our research extends the incentive mechanism literature by exploring the influence of non-persistent social status 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 short-term or long-term. Third, we use the accountability theory to interpret users' behaviors with the existence of pressure of maintaining status. It is clear that users show different behavioral change patterns compared with those in contexts where they receive badge (Anderson et al. 2013; Li et al. 2012) or incoming followers (Goes et al. 2014). The present study also yields significant implications and actionable insights for business models that rely on user contributions. It could be used to guide the incentive mechanism design with the hope of inducing more user efforts.

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 (Korfiatis et al. 2012) and lexical richness (Qiao et al. 2017). Second, we do not observe the offline behavior of these elite reviewers. The "Elite Squad" is more like a community instead of just a glory. Elite reviewers will be invited to participant some offline events (Askay and Gossett 2015). It is possible that they were invited to comment on certain types of business. Such information will help explain the underlying reason of the behavior change. 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 social status and study how such a mechanism affects user behaviors.

Acknowledgements

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

References

Anderson, A., Huttenlocher, D., Kleinberg, J., and Leskovec, J. 2013. "Steering user behavior with badges," *Proceedings of the 22nd international conference on World Wide Web - WWW '13*, pp. 95–106 (doi: 10.1145/2488388.2488398).

Ariely, D., Bracha, A., and Meier, S. 2009. "Doing Good or Doing Well? Image Motivation and Monetary Incentives in Behaving Prosocially," *American Economic Review* (99:1), pp. 544–555.

Askay, D. A., and Gossett, L. 2015. "Concealing Communities Within the Crowd: Hiding Organizational Identities and Brokering Member Identifications of the Yelp Elite Squad," *Management Communication Quarterly* (29:4), pp. 616–641 (doi: 10.1177/0893318915597301).

Baek, H., Ahn, J., and Choi, Y. 2012. "Helpfulness of Online Consumer Reviews: Readers' Objectives and Review Cues," *International Journal of Electronic Commerce* (17:2), pp. 99–126 (doi: 10.2753/JEC1086-4415170204).

Becker, S., and Ichino, A. 2002. "Estimation of Average Treatment Effects based on Propensity

Scores," *Stata Journal* (2:4), pp. 358–377 (doi: The Stata Journal).
Berger, J., Sorensen, A. T., and Rasmussen, S. J. 2010. "Positive Effects of Negative Publicity: When Negative Reviews Increase Sales," *Marketing Science* (29:5), pp. 815–827 (doi: 10.1287/mksc.1090.0557).
Bovens, M. 2010. "Two Concepts of Accountability: Accountability as a Virtue and as a Mechanism," *West European Politics* (33:5), Routledge, pp. 946–967.
Cabral, L., and Li, L. (Ivy). 2015. "A Dollar for Your Thoughts: Feedback-Conditional Rebates on eBay," *Management Science* (61:9), pp. 2052–2063 (doi: 10.1287/mnsc.2014.2074).
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.
Centola, D. 2010. "The spread of behavior in an online social network experiment.," *Science (New York, N.Y.)* (329:5996), pp. 1194–7 (doi: 10.1126/science.1185231).
Chen, Y., Harper, F. 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), pp. 1358–1398.
Chintagunta, P. K., Gopinath, S., and Venkataraman, S. 2010. "The Effects of Online User Reviews on Movie Box Office Performance: Accounting for Sequential Rollout and Aggregation Across Local Markets," *Marketing Science* (29:5), pp. 944–957 (doi: 10.1287/mksc.1100.0572).
Fandt, P. M., and Ferris, G. R. 1990. "The management of information and impressions: When employees behave opportunistically," *Organizational Behavior and Human Decision Processes* (45:1), pp. 140–158 (doi: 10.1016/0749-5978(90)90008-W).
Fayazi, A., Lee, K., Caverlee, J., and Squicciarini, A. 2015. "Uncovering Crowdsourced Manipulation of Online Reviews," *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '15*, pp. 233–242 (doi: 10.1145/2766462.2767742).
Gneezy, U., Meier, S., and Rey-Biel, P. 2011. "When and Why Incentives (Don't) Work to Modify Behavior," *Journal of Economic Perspectives* (25:4), pp. 191–210 (doi: 10.1257/jep.25.4.191).
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 (doi: 10.1287/isre.2016.0635).
Goes, P. B., Lin, M., and Yeung, C. A. 2014. "' Popularity Effect ' in User-Generated Content : Evidence from Online Product Reviews," *Information Systems Research* (25:2), pp. 222–238.
Hennig-Thurau, T., Gwinner, K. P., Walsh, G., and Gremler, D. D. 2004. "Electronic word-of-mouth via consumer-opinion platforms: What motivates consumers to articulate themselves on the Internet?," *Journal of Interactive Marketing* (18:1), pp. 38–52 (doi: 10.1002/dir.10073).
Hochwarter, W. A., Ferris, G. R., Gavin, M. B., Perrewé, P. L., Hall, A. T., and Frink, D. D. 2007. "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), pp. 226–239 (doi: 10.1016/j.obhdp.2006.09.003).
Hosanagar, K., Fleder, D., Lee, D., and Buja, A. 2014. "Will the Global Village Fracture Into Tribes? Recommender Systems and Their Effects on Consumer Fragmentation," *Management Science* (60:4), pp. 805–823.
Jabr, W., Mookerjee, R., Tan, Y., and Mookerjee, V. S. 2014. "Leveraging philanthropic behavior for customer support: the case of user support forums," *MIS Quarterly* (38:1), pp. 187–208.
Kahneman, D., and Tversky, A. 1979. "Prospect Theory: An Analysis of Decision under Risk," *Econometrica* (47:2), pp. 263–292 (doi: 10.2307/1914185).
Keegan, J., and Kabanoff, B. 2007. "Indirect Industry-and Subindustry-Level Managerial Discretion Measurement," *Organizational Research Methods* (11:4), pp. 682–694 (doi: 10.1177/1094428107308897).
Korfiatis, N., García-Bariocanal, E., and Sánchez-Alonso, S. 2012. "Evaluating content quality and helpfulness of online product reviews: The interplay of review helpfulness vs. review content," *Electronic Commerce Research and Applications* (11:3), pp. 205–217.
Lerner, J. S., and Tetlock, P. E. 1999. "Accounting for the Effects of Accountability," *Psychological Bulletin* (125:2), pp. 255–275.
Li, Z., Huang, K.-W., and Cavusoglu, H. 2012. "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)*.

Luca, M. (n.d.). "Reviews , Reputation , and Revenue: The Case of Yelp.com," *SSRN Electronic Journal*, pp. 1–40 (doi: 10.2139/ssrn.1928601).

Luca, M., and Zervas, G. 2016. "Fake It Till You Make It: Reputation, Competition, and Yelp Review Fraud," *Management Science* (April), pp. 1–16 (doi: 10.2139/ssrn.2293164).

Milinski, M., Semmann, D., and Krambeck, H.-J. 2002. "Reputation helps solve the 'tragedy of the commons,'" *Nature* (415:6870), pp. 424–426 (doi: 10.1038/415424a).

Mudambi, S. M., and Schuff, D. 2010. "What makes a helpful online review? A study of customer reviews on Amazon. com," *Management Information Systems Quarterly* (34:1), pp. 185–200.

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.

Qiao, D., Whinston, A. B., and Lee, S. 2017. "Incentive Provision and Pro-Social Behaviors," in *50th Hawaii International Conference on System Sciences*, pp. 5599–5608.

Roberts, J. A., Hann, I.-H., and Slaughter, S. A. 2006. "Understanding the Motivations, Participation, and Performance of Open Source Software Developers: A Longitudinal Study of the Apache Projects," *Management Science* (52:7), pp. 984–999 (doi: 10.1287/mnsc.1060.0554).

Rosenbaum, P. R., and Rubin, D. B. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika* (70:1), pp. 41–55 (doi: 10.2307/2335942).

Ryan, R. M., and Deci, E. L. 2000. "Intrinsic and Extrinsic Motivations: Classic Definitions and New Directions," *Contemporary Educational Psychology* (25:1), pp. 54–67 (doi: 10.1006/ceps.1999.1020).

Schlenker, B. R., Weigold, M. F., and Doherty, K. 1991. "Coping with accountability: Self-identification and evaluative reckonings," in *Handbook of Social and Clinical Psychology: The Health Perspective*, Pergamon Press.

Staw, B. M. 1976. "Knee-deep in the big muddy: a study of escalating commitment to a chosen course of action," *Organizational Behavior and Human Performance* (16:1), pp. 27–44 (doi: 10.1016/0030-5073(76)90005-2).

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

Tetlock, P. E., and Boettger, R. 1989. "Accountability: A Social Magnifier of the Dilution Effect," *Journal of Personality and Social Psychology* (57:3), pp. 388–398 (doi: 10.1037/0022-3514.57.3.388).

Vance, A., Lowry, P. B., and Eggett, D. 2013. "Using Accountability to Reduce Access Policy Violations in Information Systems," *Journal of Management Information Systems* (29:4), pp. 263–290 (doi: 10.2753/MIS0742-1222290410).

Vance, A., Lowry, P. B., and Eggett, D. 2015. "Increasing accountability through user-interface design artifacts: A new approach to addressing," *MIS Quarterly* (39:2), pp. 345–366.

Wasko, M. 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.

Willer, R. 2009. "Groups reward individual sacrifice: The status solution to the collective action problem," *American Sociological Review* (74:1), pp. 23–43 (doi: 10.1177/000312240907400102).

Williams, K., Harkins, S., and Latane, B. 1981. "Indentifiability as a Deterrent to Social Loafing: Two Cheering Experiments," *Journal of Personality and Social Psychology* (40:2), pp. 303–311 (doi: 10.1037/0022-3514.40.2.303).

Zhu, F., and Zhang, X. 2010. "Impact of Online Consumer Reviews on Sales: The Moderating Role of Product and Consumer Characteristics," *Journal of Marketing* (74:2), pp. 133–148.

Zhu, L., Yin, G., and He, W. 2014. "Is This Opinion Leader'S Review Useful? Peripheral Cues for Online Review Helpfulness," *Journal of Electronic Commerce Research* (15:4), pp. 267–280.