On the Effectiveness of Self-Contained Reward Systems to Incentivize User-Generated Content

Alexander Kupfer
University of Innsbruck
Alexander.kupfer@uibk.ac.at

Dominik Gutt
Erasmus University Rotterdam
gutt@rsm.nl

Steffen Zimmermann
University of Ulm
Steffen.zimmermann@uni-ulm.de

Dennis Kundisch
Paderborn University
dennis.kundisch@uni-paderborn.de

Abstract

Many digital platforms rely on reward systems to incentivize the production of user-generated content. These platforms frequently resort to financial or peer-based reward systems. Both systems have drawbacks and limitations. Financial rewards such as vouchers can crowd out effort of the content contributor and induce a positivity bias to the user-generated content. Peer-based rewards, which are conditional on upvotes from peers, are used when crowds identify the best contribution or curation. Because online review platforms aim at incentivizing a variety of contributions at scale, they can hardly apply peer-based reward systems. It is undetermined whether gamification systems designed for self-contained rewards—where reviewers receive points and badges for their activities unconditional on upvotes from peers—can effectively circumvent these drawbacks and limitations. To empirically address this question, we draw on a data set of online reviews from Google Maps and Tripadvisor that we matched on a site level. Our identification strategy hinged on a natural experiment of Google's self-contained reward system, Local Guides, being restructured such that particular reviewing activities were rewarded with more points. We found that self-contained reward systems avoid effort crowding out and positivity bias but do incentivize the production of user-generated content. Beyond that and in contrast to our expectation, we detected substantial positive spillover effect to the unincentivized task of submitting a rating without a textual review. Furthermore, we documented that the effectiveness of self-contained rewards differs across low- and high-expertise reviewers. Most importantly, our study shows that peer-based incentive mechanisms are no prerequisite for effective nonfinancial reward systems, and applying self-contained reward systems circumvents many negative side effects associated with financial rewards.

Keywords: Online Review Platform; Self-Contained Reward System; Reviewing Effort; Spillover Effects; Gamification; Google; Tripadvisor

1 Introduction

User-generated content platforms such as Wikipedia, StackOverflow, and Tripadvisor rely on users to provide content—be it encyclopedia entries, help with coding, or online reviews. To nurture the growth of their platforms and the revenues of their business, they incentivize users to increase the extensive (how many posts) and intensive (how long the posts) margin of their contribution. In a recent report, Accenture states that more than 90% of companies currently employ engagement programs to foster user-generated content (Accenture, 2017). Such practices have become so vital for the general economy that the Federal Trade Commission (FTC) has issued guidelines on the allowed practices of incentivizing user-generated content (FTC, 2017).

Past research has demonstrated the effects of financial rewards and peer-based rewards. Financial rewards can be vouchers, rebates, or cash transfer and have been particularly popular among sellers on e-commerce platforms. Peer-based rewards are nonfinancial and are awarded conditional on upvotes by peers and can be found, for instance, on Wikipedia, StackOverflow, and Reddit.

Studies have found that financial rewards can increase the extensive margin of contributions, such as the number of reviews (Cabral and Li, 2015; Khern-am-nuai et al., 2018; Qiao et al., 2020; Yu et al., 2020). Yet, these rewards also crowd out the intensive margin of contributions (reviewing effort in terms of review length), induce a positive bias in the valence of contributions (rating valence) (Cabral and Li, 2015; Khern-am-nuai et al., 2018; Qiao et al., 2020; Yu et al., 2022), and entail substantial marginal costs for the solicitor.

Several studies have found evidence for the effectiveness of peer-based rewards (Burtch et al., 2021; Gallus, 2017; Goes et al., 2016). This kind of reward is applied when a platform identifies the best user contribution based on peers upvotes. It is a well-established method; StackOverflow uses it to identify the best response to a question, and Wikipedia uses it to curate the entries that are most accurate. As such, peer-based rewards are used for both selecting the best responses by a crowd of contributors and motivating users to contribute.

By contrast, online review platforms have a different objective than such knowledge platforms. They are not aiming for the single best review for a tourist site, for example. Rather, they are aiming at encouraging a wide variety and scale of reviews. Tourist sites, for example, can have many attributes that can be perceived differently. It is highly unlikely that one review can capture all the facets of a site. One single review may not even be desired; a high number of reviews can indicate the appeal of a tourist site, product, or service. Peer-based reward systems aim not to incentivize a variety of ideas, but to select the best one; therefore, they are not compatible with the objectives of online review platforms.

One platform design mechanism that could circumvent the downsides of financial reward systems and the drawback that peer-based rewards may be inapplicable to platforms seeking to incentivize contributions at variety and scale is that of a self-contained reward system using gamification (Liu et al., 2017). Yet, its effectiveness on extensive and intensive margins is unstudied so far. Gamification systems are a powerful social technology that harnesses users' desire for social status, community relatedness, and opportunity to display their competence (Hamari et al., 2014). Rewards typically come in the form of points and badges and are nonfinancial in nature. Although peer-based reward also relies on gamification systems, the endowment in a self-contained reward system is independent of peer votes, or self-contained. Empirical field evidence on the effectiveness of self-contained reward systems in fostering the contribution of user-generated content is scant. Consequently, in this paper, we sought answers to the following research questions:

(1) *Can self-contained nonfinancial reward systems effectively incentivize content contribution?*

(2) *Can self-contained nonfinancial reward systems prevent effort crowding out and positivity bias?*

We empirically examined these questions using data from a natural experiment created by the restructuring of the self-contained reward system of Google Maps. We operationalized reviewing effort by using the number of textual reviews[1] to capture the extensive margin of content contribution and using the textual review length to capture the intensive margin. Textual reviews offer consumers a better understanding of different facets of a product or service. The textual review length directly correlates with the helpfulness and the level of information that reviewers provide (Cao et al., 2011; Pan and Zhang, 2011). Because review length has decreased significantly over the years, a detailed analysis of motivational aspects to stimulate the creation of long reviews is timely and relevant (Reviewtrackers, 2021). To examine a potential positivity bias, we followed existing literature and operationalized review valence with its numerical rating.

Because online review platforms aim to utilize these rewards in a targeted and effective manner, it is important to understand their potential side effects. For this purpose, we also examined whether there are spillover effects to unincentivized tasks because this might have unintended consequences on the overall quality of the content on the online review platform. The unincentivized activity in our study was represented by nontextual reviews, i.e., rating-only reviews, which are not particularly incentivized. Furthermore, we investigated whether self-contained rewards are effective for all users universally. We separately investigated users with low and high reviewing expertise to test whether there are heterogeneous effects of these rewards across users.

We guided our empirical analyses using hypotheses rooted in Self-Determination Theory (SDT) by Ryan and Deci (2000). As empirical context, we took advantage of the restructuring of the Google Maps reward system, Local Guides, which represents an ideal natural experiment. Google Maps has set up Local Guides to encourage its users with self-contained rewards to participate in reviewing activities. When users perform certain activities, they are directly rewarded with points that allow them to reach a number of successive levels over time. Between fall 2017 and spring 2018, Google changed the reward scheme of the program by increasing the incentives for textual reviews. To allow for an unconfounded

[1] In our study, the number of textual reviews was a subset of the total review quantity. On our focal platform Google Maps, reviewers did not have to use textual reviews for their ratings.

estimation, we applied a Difference-in-Difference (DiD) analysis with a control group of reviews from Tripadvisor as counterfactuals so we could eliminate confounding factors such as platform-specific trends and seasonality.

In contrast to the literature of financial rewards (Burtch et al., 2018; Cabral and Li, 2015; Khern-am-nuai et al., 2018; Qiao et al., 2020; Sun et al., 2017; Wang et al., 2016; Yu et al., 2022), we found that self-contained rewards do not exhibit effort crowding out. Both the extensive margin and the intensive margin of content contribution increased substantially, by 62% and 16%, respectively. Moreover and in contrast to our hypothesis, we found that such rewards also have positive spillover effects to unincentivized tasks. We found that unincentivized nontextual reviews increased by 82%, indicating that the spillover effect is even higher than the effect on targeted activity. Finally, our results suggest that self-contained rewards do not induce a statistically significant positivity bias into numerical average ratings, as financial rewards do (Khern-am-nuai et al., 2018).

Our results also showed consistent patterns across user groups, suggesting that the gamification system is comparably effective for low- and high-expertise users. However, consistent with predictions of SDT, we found that the incentives for writing long reviews were effective only for high-expertise users. In contrast, the spillover effects to nontextual reviews were more pronounced for low-expertise users. We conducted our primary analysis on a site level and complemented it with fine-grained user-level observations that yielded consistent results, highlighting the robustness of our findings.

Our study has important practical implications for designers and managers of user-generated content platforms. Our findings suggest that self-contained reward systems can successfully incentivize the user-generated content in the form of online reviews while preventing negative side effects associated with financial rewards and avoiding reliance on the conditionality of peer votes. However, our study also highlights how important awareness of potential side effects; the ratio of nontextual review to textual review is higher after the restructuring that focuses on the incentivization of textual reviews. Furthermore, our findings suggest that users respond differently to self-contained rewards that address their reviewing effort. Therefore, online review platforms might face a tradeoff between which user group to focus on with the introduction of self-contained rewards.

We thus make two important contributions to research with this paper: First, we contribute to the literature on user-generated content incentivization (Burtch et al., 2021; Burtch et al., 2018; Cabral and Li, 2015; Gallus, 2017; Goes et al., 2016; Khern-am-nuai et al., 2018) by providing evidence for the effectiveness of self-contained rewards in stimulating content generation and avoiding effort crowding out as well as positivity bias. Second, we contribute to the literature on gamification (Santhanam et al., 2016; Hamari et al., 2014; Kankanhalli et al., 2012) by offering empirical evidence that gamification system effectiveness is heterogeneous in its effects across user groups and regarding spillovers to unincentivized tasks. These insights can yield valuable insights for gamification systems designers that recent literature has called for (Schöbel et al., 2020).

2 Related Literature

Extant literature has examined various aspects of online reviews (Babić Rosario et al., 2016; Floyd et al., 2014; Gutt et al., 2019; King et al., 2014; DeMatos and Rossi, 2008). Most relevant to our work are studies on reviewer reward systems that aim to stimulate the production of reviews. Research conceptually differentiates between financial and nonfinancial rewards. Financial rewards have been shown to increase review quantity (Burtch et al., 2018; Cabral and Li, 2015; Khern-am-nuai et al., 2018; Qiao et al., 2020; Sun et al., 2017; Wang et al., 2016; Yu et al., 2022), but the effort invested for lengthy reviews declines (Khern-am-nuai et al., 2018) or remains unchanged (Stephen et al., 2012; Wang et al., 2012). Financial rewards tend to induce a positivity bias in ratings that may hamper consumer decision-making (Khern-am-nuai et al., 2018). Financial rewards can also produce negative spillover effects to other activities; for example, they may reduce reviewing effort for products whose reviews are not incentivized (Qiao et al., 2020). To alleviate these negative externalities to reviewing effort, performance-contingence can be imposed on financial incentives (Yu et al., 2022) or financial incentives can be combined with social norms (Burtch et al., 2018). Finally, studies have demonstrated that social connectedness within an online community moderates the effect of financial rewards on review quantity and reviewing effort (Sun et al., 2017).

Gamification elements are typical nonfinancial rewards (Cheong et al., 2013; Hamari et al., 2014). Rewards can consist of points that help to reach hierarchy levels, reward a badge, or combine the two (receiving a badge when a hierarchy level is reached). Users typically obtain rewards either on their own or from votes of their peers. As an example, a self-contained reward system awards users with points for each post they make, regardless of whether the other users upvote their posts. In a peer-based reward system, posting users only receive points when other users upvote them and they reach a certain threshold.

Research on the effect of nonfinancial rewards for online reviews is scarce, but such rewards have been studied in more general contexts of user-generated content, such as Wikipedia, StackOverflow, and Reddit contributions. Most prominently, randomized field experiments were applied to investigate how badges received by peer feedback impact contribution activity on Wikipedia (Gallus, 2017) and content creation on Reddit (Burtch et al., 2021). Studies found that peer-based awards increase the extensive (Gallus, 2017), respectively intensive margin (Burtch et al., 2021), of content contribution and content novelty (Burtch et al., 2021). Goes et al. (2016) investigated the effect of reward levels in an IT-related Q&A community and found that user contribution quantity increases before reaching a certain level but sharply drops afterward. All three studies examined peer-based rewards. Besides the differences in rewards' conditionality, these studies examined rewards' online communities whose aim is to filter out and motivate the best response on StackOverflow or the best-curated entry on Wikipedia. By contrast, online review systems do not seek to incentivize "the single best" review for a product or service; instead, they aim to collect a large number of varied reviews (Khern-am-nuai et al., 2022).

Therefore, even though several online review platforms, including Tripadvisor and Google Maps, have implemented self-contained reward systems, we are only aware of a few studies that have analyzed such systems. Wang and Sanders (2019) experimentally investigated the effect of self-contained financial and nonfinancial rewards. While the authors observed positive effects of both types of rewards in an online lab, it remains unknown whether the findings can also be observed in the field. Moro et al. (2019) used observational data to investigate the relationship between badges and nonfinancial rewards on Tripadvisor. The authors' exploratory prediction analysis indicates that self-contained rewards in a review system with gamification elements can predict review length and review sentiment. However, this does not answer the question of whether these gamification elements actually spur content contribution.

To the best of our knowledge, no study has yet investigated whether self-contained rewards are effective in stimulating reviewing behavior in the field.[2] The recent call for research by Qiao et al. (2020) on nonfinancial rewards in online review platforms further highlights the need for a clear understanding of the underlying effects.

It is essential to understand the effects of rewards not only on the incentivized task but also on users' unincentivized activities. As highlighted earlier, financial rewards for reviews negatively affect reviewing behavior for products whose reviews are not incentivized (Qiao et al., 2020). On knowledge-sharing platforms, however, financial rewards for contributions have been shown to positively affect unincentivized knowledge sharing (Kuang et al., 2019). The reason for these contradictory findings could lie in the way financial rewards address users' intrinsic motivation: If financial rewards address intrinsic motivation, positive spillover effects on unincentivized activities are likely, as documented by Kuang et al. (2019). Therefore, it is crucial to examine not only the effects of self-contained nonfinancial rewards in online review platforms on the incentivized task, but potential spillover effects on unincentivized activities.

Similar to spillover effects, the awareness of user heterogeneity is an increasing concern for online review platforms (Wang et al., 2019; Sun et al., 2017). While Wang et al. (2019) observe that user heterogeneity moderates user review quantity and reviewing effort when the user population is suddenly expanded, Sun et al. (2017) explicitly examine whether user heterogeneity moderates the effect of financial rewards on contribution. With this study, we aim to extend the understanding of user heterogeneity for nonfinancial rewards and investigate whether they are similarly effective for reviewers with low- and high-reviewing expertise.

Table 1 categorizes the related literature discussed along the dimensions of rewards (nonfinancial/financial), dependent variables (contribution/perception of reviews), the consideration of other effects (spillover/user heterogeneity), and underlying data (observational/experimental). It

[2] The recent paper by Wang et al. (2020) focuses on the perception of badges rather than on review contribution. The contributors found that reviews by badge holders are perceived as more competent, useful, and persuasive.

highlights our contribution to examine the effects of self-contained nonfinancial rewards on reviewing effort and taking potential spillover as well as moderating effects into account.

Table 1: Related Empirical Literature

Authors	Rewards		Focus on Review		Spill-over	User Heterog.	Data
	Nonfinancial	Financial	Contribution	Perception			
Our study	**x (sc)**		**x (qua, tex, len)**		**x**	**x**	**obs**
Moro et al., 2019	x (both)		x (len)				obs
Wang et al., 2020	x (pb)			x (help, comp)			exp
Wang and Sanders, 2019	x (sc)		x (qua, len, val)				exp
Burtch et al., 2018		x (soc)	x (qua, len)				exp
Cabral and Li, 2015		x	x (qua, val)				exp
Khern-am-nuai et al., 2018		x	x (qua, val)				obs
Qiao et al., 2020		x	x (len, val)	x (help, sent)	x		obs
Stephen et al., 2012		x		x (help, trust)			exp
Sun et al., 2017		x	x (qua)			x	obs
Wang et al., 2012		x (perf)	x (len)	x (help)			exp
Wang et al., 2016		x	x (qua, val)	x (help)			obs
Yu et al., 2022		x (perf)	x (qua, len, val)	x (help)			obs
Kuang et al., 2019		x (perf)	Contributions on Q&A platform		x		obs
Burtch et al., 2021	x (pb)		Content creation on Reddit				obs
Gallus, 2017	x (pb)		Contributions on Wikipedia				obs
Goes et al., 2016	x (pb)		Contributions on IT-Q&A Platform				obs

Abbreviations (in order of appearance): sc = self-contained, qua = review/rating quantity, tex = textual reviews, len = review length, obs = observational data, both = peer-based and self-contained, pb = peer-based, help = helpfulness, comp = competence (of the reviewer), val = valence, soc = social norms, exp = experimental data, sent = sentiment, trust = trustworthiness, perf = performance-contingent.

3 Theoretical Background and Hypotheses Development

To guide our empirical analysis, we employed SDT by Ryan and Deci (2000) as the underlying theoretical basis. SDT serves as a theoretical grounding for the effects of rewards in psychology. It also finds widespread adoption in the context of user-generated content (Burtch et al., 2018; Lou et al., 2013; Stock et al., 2015). By adopting SDT as our theory basis, we followed the recommendation of prior literature (Liu et al., 2017) to employ this theory for the investigation of user-generated content contribution behavior in gamification contexts.

SDT defines three basic human needs: competence, relatedness, and autonomy. If a reward successfully addresses these needs, individuals develop intrinsic motivation regarding a certain task (Ryan and Deci, 2000). For our hypotheses, we specifically draw on cognitive evaluation theory (CET) as a subtheory of SDT, which focuses on the impact of extrinsic rewards. Although CET has been used for financial rewards (Reimer and Benkenstein, 2016; Wang et al., 2012), it suits our context because gamification elements represent extrinsic rewards as well (Hamari et al., 2014; Kankanhalli et al., 2012;

von Rechenberg et al., 2016). According to CET, extrinsic rewards can be internalized and can positively affect intrinsic motivation if the external motivation is introduced in an informational way by targeting recipients' competence (Przybylski et al., 2010; Ryan and Deci, 2002; Wang et al., 2008).

We argue that incentivizing textual reviews with self-contained rewards represents a competence-addressing element: The more complex task of writing a review is valued substantially higher than the simple task of submitting a star rating. Thus, because the review system rewards a greater number of points to the task requiring more effort, it positively addresses reviewers' feelings of perceived competence. Naturally, a reviewer should feel more competent upon receiving more points. Furthermore, reviewers can freely choose to submit a star rating only or to additionally write an optional textual review. The rewards for textual reviews address individuals' need for autonomy (Ryan and Deci, 2000). Hence, we expect that extrinsic rewards are internalized and positively affect reviewers' motivation.[3] Our first hypothesis therefore reads as follows:

Hypothesis 1: Incentivizing textual reviews with self-contained rewards leads to more reviewing effort.

Although the rewards for nontextual reviews remain unchanged in our research environment, we still expect an indirect consequence of incentivizing textual reviews. We expect that reviewers who tend to leave only a rating without a textual review might start producing textual reviews because of the additional rewards that would motivate them—a direct consequence of Hypothesis 1. Furthermore, we have no reason to expect that users who tend to produce no review at all or users who already produce unincentivized textual reviews would start producing nontextual reviews because of the introduction of self-contained rewards for textual reviews. Therefore, we expect a negative spillover effect of self-contained rewards for textual reviews on the number of nontextual reviews. In other words, we conjecture the indirect effect on the number of nontextual reviews to be negative and state our second hypothesis as follows:

Hypothesis 2: Incentivizing textual reviews with self-contained rewards leads to a negative spillover effect on the number of nontextual reviews.

SDT refers not only to competence but also to relatedness and autonomy as basic psychological needs that are essential in promoting intrinsic motivation. Although we already outlined that incentives for textual reviews represent a competence-addressing element, we address relatedness as another psychological need. Individuals experience relatedness when they compete, cooperate, and are

[3] Note that the goal setting theory (Locke and Latham, 1990) could complement our theoretical foundation. Prior empirical literature observed that people consider reward levels as goals (von Rechenberg et al., 2016). A broad stream of literature concludes that goals direct attention to goal-relevant activities (Heath et al., 1999; Locke and Latham, 1990). Successful goal achievement displayed by reward levels is desirable because it symbolizes and increases one's competence and status (Dreze and Nunes, 2009; Lazarus, 1991; Smith, 2002). With the possibility of receiving more points for the same review task, reviewers can reach the next reward level more quickly. In this way, the relative importance of a textual review for reaching the next reward level has increased, regardless of the current reward level of the reviewer. Thus, according to goal setting theory, we would also expect the incentivized review activity to become more prevalent among reviewers because of the increased relevance for achieving the next reward level.

connected with the community itself or with other participants (Baumeister and Leary, 1995). Hence, introducing rewards for activities that aim to increase the value of the community (that is, more textual reviews) will particularly affect reviewers who feel connected with the community.[4] Consequently, we hypothesize that users with a high reviewing expertise (Wang et al., 2019) perceive incentives for textual reviews as more beneficial, resulting in an intensified effect. Therefore, we state our third hypothesis as follows:

Hypothesis 3: Incentivizing textual reviews with self-contained rewards has a stronger effect on users with high reviewing expertise.

Notably, one could argue that autonomy—the third basic psychological need according to SDT—is reduced with incentives for textual reviews because users might feel obliged to write them. In fact, the introduction of external rewards frequently coincides with motivational crowding out, meaning that intrinsic motivation is negatively influenced by the introduction of rewards (Deci and Ryan, 1985). Studies have found that when rewards impair the self-determination of people by pressuring them to think, feel, or behave in an externally controlled way, motivation is negatively affected (Deci and Ryan, 1985; Liu and Feng, 2021). In this context, Khern-am-nuai et al. (2018) as well as Burtch et al. (2018), for instance, observed that the introduction of financial rewards in an online review system leads to less reviewing effort.[5] However, we do not expect motivational crowding out for self-contained rewards because they are of gamified nature; gamification elements are generally known to stimulate intrinsic motivation and do not involve motivational crowding out (Festré and Garrouste, 2014; Liu et al., 2017; Lou et al., 2013). Furthermore, there are even more reasons to discount a motivational crowding out effect. First, the rewards for textual reviews are like a free give-away; reviewers can freely choose to submit a star rating only or to additionally write an optional textual review. Therefore, reviewers do not perceive the rewards for textual reviews as controlling and still must decide on their own how much effort to invest. Second, the introduced rewards explicitly focus on effort-related tasks, stimulating, as outlined earlier, the reviewers' feelings of perceived competence. Finally, the reviewing task itself is voluntary and, as we will outline in the next section, we collected data from public points of interest that people had most likely visited for self-determined leisure-related activities. This is in stark contrast to the concern that people perceive rewards as autonomy depriving and controlling.

[4] In general, several studies observe that users with a high reviewing expertise write more useful reviews (Baek et al., 2012; Ghose and Ipeirotis, 2011; Racherla and Friske, 2012). Although this observation is independent of rewards, it indicates that these users seem to be more concerned with the review system—being in line with SDT.

[5] Notably, Sun et al. (2017) observe that contributions from socially connected members considerably decline. This is in stark contrast to Hypothesis 3, but it is important to keep in mind that Sun et al. (2017) examined the effects of financial rewards.

4 Research Environment

We used Google Maps as our research environment to answer our research questions. As outlined before, Google unconditionally rewards Google Maps reviewers within its nonfinancial reward system, Local Guides, introduced in November 2015. At the time of its introduction, reviewers received one point for a mandatory star rating, and for each additional, but optional, review component, such as a textual review, uploading of a photo, or updating of information, they received an additional point. Prior to our observation period, other changes were simultaneously introduced: more levels to users (from 5 to 10 levels), and more points for textual reviews or uploading photos (from 1 to 5 points). If both the available levels and the points were changed, disentangling their effects would be impossible. To circumvent this problem, we focused on two changes between November 2017 and April 2018. On November 22, 2017, Google Maps introduced 5 extra points for textual reviews that were longer than 200 characters. In April 2018, the incentivization for detailing online reviews additionally increased from 5 to 10 points for both writing a textual review and writing long reviews. Thus, whereas a review with more than 200 characters was worth 6 points (1 point for the star rating plus 5 points for the textual review) prior to November 2017, it has been worth 21 points (1 point for the star rating, 10 points for the textual reviews, plus 10 extra points for writing more than 200 characters) since April 2018. Because (i) both changes occurred within a short interval, (ii) it might take some time for users to recognize the changes, and (iii) both changes incentivized textual reviews by rewarding more points, we did not analyze the time between both changes but rather the time before both changes and the time after both changes. To address the potential issue of seasonal effects, we focused on the same period but one year later (July 2017 through October 2017 compared to July 2018 through October 2018). Figure 1 summarizes the changes in the reward system and the relevant period of review data we used in our analysis.

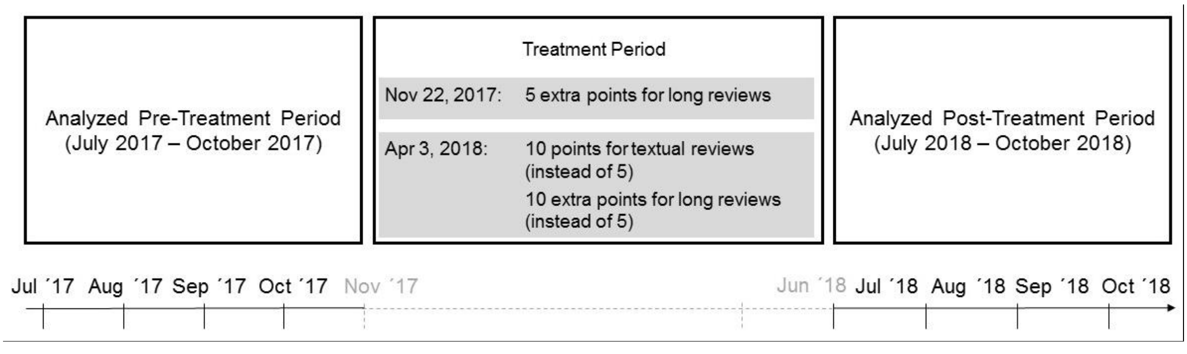
Analyzed Pre-Treatment Period
(July 2017 – October 2017)
Treatment Period
Nov 22, 2017: 5 extra points for long reviews
Apr 3, 2018: 10 points for textual reviews (instead of 5)
10 extra points for long reviews (instead of 5)
Analyzed Post-Treatment Period
(July 2018 – October 2018)
Jul ´17
Aug ´17
Sep ´17
Oct ´17
Nov ´17
Jun ´18
Jul ´18
Aug ´18
Sep ´18
Oct ´18

Figure 1. Sequence of change in Google's reward system and relevant data for analysis

Locations that are typically reviewed on Google Maps include bars, restaurants, hotels, and shops. These locations are subject to time variability (for example, increases in prices or a new restaurant chef) and would be problematic for this analysis if the variability affected the reviewing effort. Therefore, we focused on locations that were less sensitive to time variability. We further set the conditions that locations were appropriate for the analysis only if they (i) did not charge visitors, (ii) were accessible to the public, and (iii) were reviewed on Google Maps. We decided in favor of points of interest and selected bridges, squares, fountains, and monuments as sites that matched all our requirements. For each site group (bridges, squares, fountains, and monuments), we selected 10 points of interest within Europe. Our final dataset included 40 sites, ranging from the 25 de Abril Bridge in Lisbon to the Freedom Monument in Latvia. Review dates on Google Maps did not have a distinct timestamp but a relative date, such as "one month ago." (Relative dates on a monthly basis are available for one year only; after that, the relative dates are shown only in a yearly frequency, such as "a year ago.") In early November 2017 and early November 2018, we used web scraping tools to extract all existing reviews that were written during the previous four months. This gave us the relevant review information on a monthly frequency for both periods. We excluded three sites because of inconsistent data; data cannot be downloaded retrospectively, so others could not substitute these sites. Consequently, we had monthly review data for the same time of the year before and after adapting the reward system for 37 sites across Europe. For each review, we retrieved the relative date, the star rating, and the textual review, if it was available.

Table 2 presents the main variables of our analysis on site level: To operationalize reviewing effort, we used the number of textual reviews ($TEXREV$) and the review length measured in number of characters ($REVLENGTH$). The number of nontextual reviews ($NONTEXREV$) represents the unincentivized task. Finally, we use the average rating (AVG_RATING) to examine whether self-contained reward systems induce a positivity bias. In Panel A of Table 2, we present descriptive statistics of reviews for each site group individually and for all sites in aggregate that we used in the paper's analysis. Panel B and Panel C of Table 2 further differentiate between the Pre-Treatment Period and the Post-Treatment Period, respectively.

Table 2: Descriptive Statistics of Extracted Google Maps Reviews

Panel A: Entire Dataset

	All Sites	Monument	Squares	Fountains	Bridges
AVG_RATING	4.45	4.49	4.37	4.46	4.49
NON_TEXREV	31,740	7,880	9,441	6,169	8,250
TEXREV	17,051	4,267	4,801	3,422	4,561
REVLENGTH	73	79	69	78	68

Panel B: Pre-Treatment Period

	All Sites	Monument	Squares	Fountains	Bridges
AVG_RATING	4.42	4.46	4.36	4.44	4.46
NON_TEXREV	12,696	2,431	4,450	1,801	4,014
TEXREV	7,008	1,387	2,315	1,116	2,190
REVLENGTH	63	67	60	66	62

Panel C: Post-Treatment Period

	All Sites	Monument	Squares	Fountains	Bridges
AVG_RATING	4.47	4.50	4.38	4.47	4.51
NON_TEXREV	19,044	5,449	4,991	4,368	4,236
TEXREV	10,043	2,880	2,486	2,306	2,371
REVLENGTH	80	85	77	84	74

If we used only Google data, however, myriad confounding factors, such as platform-specific trends, could obscure the causal effect of the reward system change on our dependent variables of interest. In particular, Google Maps could benefit from a high inflow of users regardless of the change, or the post-treatment period could have stronger tourism figures than the pre-treatment period. To rule out these potentially confounding factors and be able to analyze the casual effect of the reward system change on reviewer behavior, we needed to identify counterfactuals for our sites of interest. As a result, we were able to analyze what would have happened to the number of textual reviews, textual review length, nontextual reviews, and average rating of a particular site in the absence of the reward system change. To this end, we additionally drew on reviews from Tripadvisor as counterfactuals to compare the development over time on both online review platforms. For this purpose, we collected all reviews on Tripadvisor for the same sites using a self-developed web scraper. In Tripadvisor, writing a textual review is mandatory; therefore, the number of textual reviews was equal to total review quantity for that platform. Even though there are no nontextual Tripadvisor reviews, the Tripadvisor reviews are valid counterfactuals for nontextual Google Maps reviews. The counterfactual reviews on Tripadvisor create a framework that rules out confounding factors such as platform-specific trends or seasonality. As long as both textual and nontextual reviews are equally prone to these factors, textual reviews on TripAdvisor can be valid counterfactuals for nontextual reviews on Google Maps. This can be tested with the parallel trends assumption; if the parallel trends assumption holds, the counterfactuals are valid. Finally, it is important to note that Tripadvisor's reward system, TripCollective, did not change during our observation period.

5 Empirical Analysis

Figure 2 shows the number of textual reviews ($ln(TEXREV)$), their average length ($ln(REVLENGTH)$), the number of nontextual reviews ($ln(NONTEXREV)$), and the average rating (AVG_RATING) for the 37 sites. By comparing the pre-treatment period (July 2017 through October 2017) with the post-treatment period (July 2018 through October 2018), we see that the number of textual reviews, their average length, and the number of nontextual reviews exhibit a distinct increase during the post-treatment period (July 2018 through October 2018). Moreover, Google Maps and Tripadvisor exhibited parallel trends before the treatment. We formally test the parallel trends assumption with a lags and leads model in section 5.5.4 and, reassuringly, find that the parallel trends assumption holds.

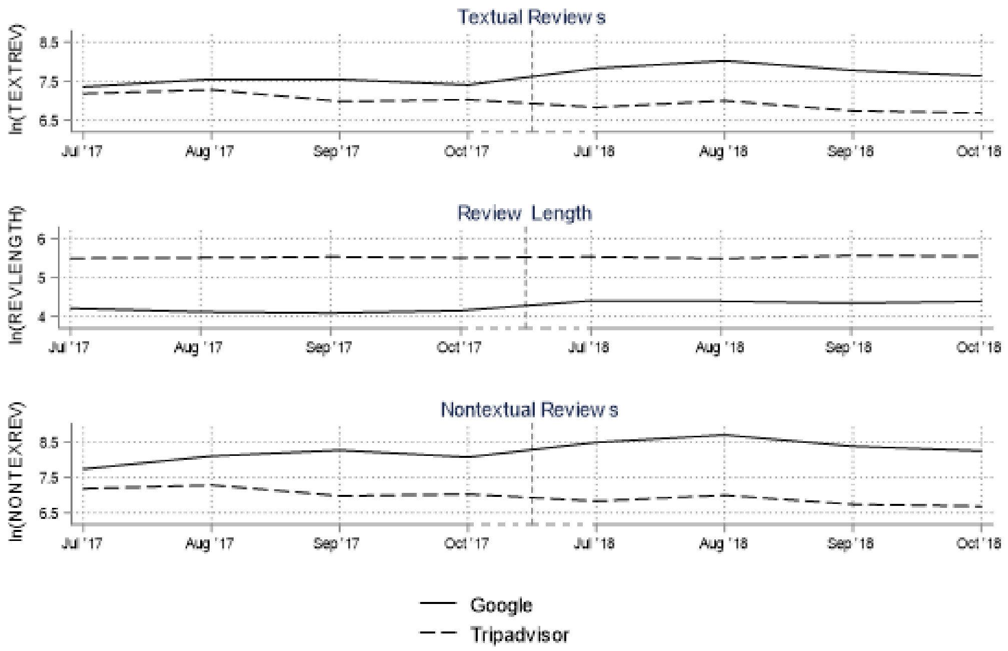

Figure 2. Textual Reviews, Review Length, and Nontextual Reviews of Tripadvisor and Google Maps Around the Reward System Restructuring

The basic idea of this approach was to compare two different samples: one receiving a treatment (here, Google Maps, whose reward system was changed) and one not (here, Tripadvisor, whose reward system was unchanged). Hence, we refer to Google Maps as the treatment group and to Tripadvisor as the control group. To test Hypothesis 1, we used $TEXTREV$ and $REVLENGTH$. To ensure the comparability of review lengths even when they were written in different languages, we translated all non-English textual reviews to English using Google Translate. To test Hypothesis 2, we used $NONTEXREV$. For Hypothesis 3, we conducted the analysis from Hypotheses 1 and 2 separately for low- and high-expertise reviewers, which we identified using the number of reviews they had written. In the following, we provide a detailed outline of our econometric model for Hypotheses 1 and 2 separately.

5.1 Reviewing Effort (H1)

Number of Textual Reviews. To gauge how the number of textual reviews would change in response to the reward system restructuring, we estimated a standard DiD model. For this model, we restricted our estimation sample to reviews that had review text. Formally, in the standard DiD model, we defined (i) a dummy variable $GOOGLE$ being 1 if the dependent variable was Google Maps data (the treatment group) and being 0 for Tripadvisor data and (ii) a $TIME$ dummy being 1 if the dependent variable was in the post-treatment period and being 0 otherwise. The regression model included both dummies and an interaction term of both dummies $GOOGLE \times TIME$. We further included site-level fixed effects μ_s to control for site-specific unobserved heterogeneity. Finally, we logarithmized our dependent variable (monthly number of textual reviews) to fit the assumptions of the ordinary least squares regression, and we applied robust standard errors clustered at the site level. The regression model can be summarized as follows:

$$\ln(TEXREV_{it}) = \alpha + \beta_1 GOOGLE_i + \beta_2 TIME_t + \beta_3 (GOOGLE_i \times TIME_t) + \mu_s + \varepsilon_{its} \qquad (1)$$

where $\ln(TEXREV_{it})$ is the log-transformed number of textual reviews for site *i* in month *t*.

The treatment effects $(GOOGLE_i \times TIME_t)$ for the post-treatment period are significant at the 0.1% level and have a value of 0.619. This coefficient implies that the change in the reward system increased the number of textual reviews by about 62%. The results of the regression coefficients are presented in Column (1) of Table 3. Thus, with reviewing effort proxied by the number of textual reviews, we found support for H1.

Table 3: Main Results

	ln(TEXREV)	*ln(REVLENGTH)*	*ln(NONTEXREV)*
	(1)	(2)	(3)
GOOGLE x TIME	0.619*** (0.121)	0.165*** (0.042)	0.822*** (0.139)
Site-Level FE	✓	✓	✓
Observations	586	586	586
Adj. R^2	0.65	0.89	0.67

Note: Robust standard errors are in parentheses. $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$.

Textual Review Length. To analyze our second proxy for reviewing effort (whether the change in the reward system increased the textual review length), we estimated the same DiD model using textual review data from Google Maps and Tripadvisor. Although a minimum number of characters is mandatory on Tripadvisor, the DiD analysis could identify the treatment effect because our analysis focused on the changes in textual review length on Google Maps, and the Tripadvisor data solely functioned as the control group. The regression model can be summarized as follows:

$$\ln(REVLENGTH_{it}) = \alpha + \beta_1 GOOGLE_i + \beta_2 TIME_t + \beta_3 (GOOGLE_i \times TIME_t) + \mu_s + \varepsilon_{its} \qquad (2)$$

where $\ln(REVLENGTH_{it})$ is the log-transformed average textual review length (in number of characters) for site i in month *t*.

The treatment effects $GOOGLE_i \times TIME_t$ for the post-treatment period are significant at the 0.1% level and have a value of 0.165. Because we used the logarithm of the average textual review length, this coefficient implies that the change in the reward system increased the average textual review length by 17%. The results of the regression coefficients are presented in Column (2) of Table 3. Therefore, we also find support for Hypothesis 1 when using the length of textual reviews as proxy for reviewing effort. Importantly, we do not observe motivational crowding out, suggesting that explicit rewards for textual reviews are internalized and positively affect reviewers' motivation.

5.2 Spillover to Nontextual Reviews (H2)

We then investigated whether incentivizing textual reviews causes a negative spillover effect on the number of nontextual reviews. For this purpose, we used the number of nontextual reviews from Google Maps and, as before, the number of reviews from Tripadvisor. Our regression model to test Hypothesis 2 was denoted as follows:

$$\ln(NONTEXREV_{it}) = \alpha + \beta_1 GOOGLE_i + \beta_2 TIME_t + \beta_3 (GOOGLE_i \times TIME_t) + \mu_s + \varepsilon_{its} \qquad (3)$$

where $\ln(NONTEXREV_{it})$ is the natural logarithm of the number of nontextual reviews for site i in month t.

The treatment effects $GOOGLE_i \times TIME_t$ for the post-treatment period were significant at the 0.1% level and had a value of 0.822 (Column (3) of Table 3). This coefficient implies that the change in the reward system increased the number of nontextual reviews by about 82%.

This result contradicts Hypothesis 2. While we expected the spillover effect to be negative, we observed a substantial positive spillover effect on the number of nontextual reviews due to the incentivization of textual reviews. Comparing the coefficients in Column (3) of Table 3 with the coefficient in Column (1) of Table 3 indicates that incentivizing textual reviews has in fact a higher spillover effect on the number of nontextual reviews than on the intended outcome of more textual reviews. More generally, this finding indicates that financial rewards can lead to spillover effects (as documented by Qiao et al., 2020), but also that self-contained rewards directly affect other unincentivized activities.

5.3 Moderating Effect of Reviewing Expertise (H3)

Our Hypothesis 3 states that incentivizing textual reviews has a stronger effect on users with high reviewing expertise. To analyze this, we used the number of reviews a user had written to classify users' reviewing expertise. Importantly, this requires the number of reviews written by a user for both the treatment group and the control group, respectively. We split our sample along the group-specific median of this variable to create one dataset with users who have a low reviewing expertise and one dataset with users who have a high reviewing expertise. We then re-estimated the models in Equations (1) to (3) on these two subsamples. The results are presented in Table 4.

Essentially, we observed that the number of textual reviews increased for users with both low and high reviewing expertise, suggesting that the restructuring appealed to both groups. Yet, only users with high reviewing expertise exhibited an increase in textual review length upon the system restructuring. Although this only partially supports Hypothesis 3 stating that incentivizing textual reviews has a stronger appeal for users with high reviewing expertise, it is in line with the SDT: Rewards for the more effortful task (writing long reviews) are only effective for high-expertise users. Finally, we can see that the system restructuring had a more positive spillover effect on the number of nontextual reviews of low-expertise users than of high-expertise users.

Table 4: Main Results for Low/High Reviewing Expertise

	ln(TEXREV)		*ln(REVLENGTH)*		*ln(NONTEXREV)*	
	(1) Low Expertise	(2) High Expertise	(3) Low Expertise	(4) High Expertise	(5) Low Expertise	(6) High Expertise
GOOGLE x TIME	0.537*** (0.159)	0.469*** (0.115)	0.047 (0.093)	0.236*** (0.049)	0.791*** (0.169)	0.646*** (0.131)
Site-Level FE	✓	✓	✓	✓	✓	✓
Observations	445	528	445	528	452	526
Adj. R²	0.489	0.709	0.734	0.867	0.674	0.592

Note: Robust standard errors are in parentheses. *p < 0.05; ** p < 0.01; *** p < 0.001.

5.4 Further Analysis of the Average Rating

We wanted to analyze whether the reward system restructuring had an effect on the average rating of the reviews. Past research has consistently found positivity biases in the numerical rating in response to financial rewards (Burtch et al., 2018; Cabral and Li, 2015; Khern-am-nuai et al., 2018; Qiao et al., 2020; Sun et al., 2017; Wang et al., 2016; Yu et al., 2022). Presenting biased information to users of third-party platforms is undesirable. Positively biased ratings might impair a user's decision-making on the platform and lead to distrust and discontinuation of using the platform (Filippas et al., 2022). To investigate this, we computed the average rating at the site level for all Google Maps and Tripadvisor ratings and named the variable AVG_RATING. We inserted AVG_RATING as the dependent variable into our main DiD regression framework to investigate whether the restructuring positively biased the average ratings.

Table 5 displays the results. We found that the restructuring induced no positivity bias into the Google ratings. In fact, the coefficient of $GOOGLE_i \times TIME_t$ was not statistically significant at any reasonable level.

Table 5: Analysis of Positivity Bias in Ratings

	AVG_RATING
	(1)
GOOGLE x TIME	–0.010 (0.045)
Site-Level FE	✓
Observations	586
Adj. R²	0.40

Note: Robust standard errors are in parentheses. $^{*}p$ < 0.05; ** p < 0.01; *** p < 0.001.

Based on the evidence from our counterfactual analysis with Tripadvisor, we conclude that (i) the restructuring provided an effective stimulus for users to increase reviewing effort, (ii) there is an unexpected positive spillover effect to the number of nontextual reviews, (iii) in terms of textual review length, users with high reviewing expertise increase their reviewing effort more than users with low reviewing expertise, and (iv) the average ratings have no positivity bias.

5.5 Robustness Checks

In this section, we discuss the robustness of our results.

5.5.1 Individual-Level Analysis

Until now, we have presented evidence of the effectiveness of the Google Maps gamification restructuring on the aggregate site level. This has enabled us to construct a counterfactual group using the same sites on Tripadvisor to tease out the causal effect of the restructuring. To coincide with drawing upon individual-level CET, we present additional empirical evidence on the user level. We collected the entire reviewing history of a new randomized sample of Google Maps reviewers. To increase the generalizability of our study beyond observations from the EU and beyond free tourist sites, we used four points of interest at the Las Vegas Strip as "initial seed" and selected their latest 1,000 reviewers. The sites included the Welcome to Fabulous Las Vegas sign, Fountains of Bellagio, and two In-N-Out Burger restaurants. We scraped each user's reviewing history to collect all the reviews and ratings for all locations these users had visited. We used this data in the following analysis. Similar to our prior analysis, we had to deal with the Google Maps relative time stamps for each review. All reviews older than one year were tagged with yearly relative time stamps only ("a year ago," "two years ago," and so on). Therefore, a one-time scrape of the user history in the beginning of November 2018 did not provide the relevant information for this analysis because we could not exactly determine on what date reviews were written before November 2017. We consequently rescraped the review history of the selected users in the beginning of July 2019 so that we could identify the reviews that were written between July 2017 and October 2017.[6] Date extraction for our second period (July 2018 to October 2018) was more convenient because we obtained the monthly frequency for the beginning of November 2018 scrape anyway. Our final dataset included more than 41,000 reviews of 1,166 individual users and covered the same periods as in our main DiD analysis (pre-treatment: July 2017 to October 2017; post-treatment: July 2018 to October 2018).

Similar to our main DiD analysis, we calculated pre-treatment period and post-treatment period averages for (i) the number of textual reviews, (ii) textual review length, and (iii) the number of nontextual reviews for each user. Our main variable of interest was $TIME$, which was 0 for reviews

[6] To be specific, all reviews written between July 2017 and November 2017 had the timestamp "a year ago" for both the beginning of November 2018 and the beginning of July 2019 scrape. Reviews with a timestamp of "a year ago" in November 2018 but "two years ago" in July 2019 were written before July 2017.

given before the restructuring and 1 after. To mitigate concerns of time constant-user level factors that might confound our analysis, we implemented user-level fixed effects. This analysis did not employ a DiD with counterfactuals because it was technically impossible to find matching users on Tripadvisor without access to proprietary Tripadvisor data. Nevertheless, this analysis reliably detected increases in reviewing behavior on Google Maps after the restructuring. The results of this analysis, presented in Table 6, are consistent with our main analysis in Table 3. We found additional support for Hypothesis 1 (Column (1)-(2)). As before, we also observed a significant and positive spillover effect on the number of nontextual reviews, which contradicts Hypothesis 2.

Table 6: User-Level Analysis

	ln(TEXREV)	*ln(REVLENGTH)*	*ln(NONTEXREV)*
	(1)	(2)	(3)
TIME	0.339*** (0.047)	0.077** (0.025)	0.410*** (0.050)
User-Level Fixed Effects	✓	✓	✓
Observations	1,684	1,684	1,454
Adj. R^2	0.704	0.770	0.719

Note: Robust standard errors are in parentheses. *p < 0.05; ** p < 0.01; *** p < 0.001.

We again differentiated between users with low and high reviewing expertise based on the sample's median value for review submissions and re-estimated Equations (1) to (3). Remarkably, the results (shown in Table 7) were highly consistent with our results from the prior analysis shown in Table 4 and therefore partially support Hypothesis 3. Even the stronger response in nontextual reviews by low-expertise users compared to high-expertise users was reflected in the user-level analysis. These results also confirm that the stronger increase in review length for high-expertise reviewers, respectively the stronger increase in nontextual reviews for low-expertise reviewers, are not driven merely by a reviewer's unobservable characteristics because we control for them using user-level fixed effects.

Table 7: Analysis for Different User Expertise Levels

	ln(TEXREV)		*ln(REVLENGTH)*		*ln(NONTEXREV)*	
	(1) Low Expertise	(2) High Expertise	(3) Low Expertise	(4) High Expertise	(5) Low Expertise	(6) High Expertise
TIME	0.330*** (0.064)	0.344*** (0.063)	–0.015 (0.050)	0.127*** (0.028)	0.514*** (0.075)	0.302*** (0.066)
User-Level Fixed Effects	✓	✓	✓	✓	✓	✓
Observations	602	1,082	602	1,082	746	708
Adj. R^2	0.533	0.603	0.746	0.790	0.686	0.760

Note: Robust standard errors are in parentheses. *p < 0.05; ** p < 0.01; *** p < 0.001.

5.5.2 Rebates and Perks on Google Maps

The self-contained reward system is not the only way Google incentivizes content contribution. We will discuss the other methods and explain why they are unlikely to confound our estimates. The most prominent concerns are (i) quasi-monetary incentives for Google Maps reviewers (such as Google Play Music rebates, free Google Drive Storage, or other perks and discounts) and (ii) Google Maps push notifications on Android smartphones asking for reviews that may affect the reviewing behavior. Yet, based on the anecdotal evidence we collected from news outlets and user forums, these two concerns were not perfectly and temporally correlated with our treatment. Google Maps notifications existed before our treatment, and there were no updates related to notifications in the Google Maps apps for Android (APKMirror, 2020) or iOS (iPa4Fun 2020), respectively. Regarding perks and rebates, there is no automatic mechanism awarding users a specific offer in advance. Further, Google has offered rebates and perks on an ongoing but irregular basis, and only for a small number of reviewers. These offers are unknown in advance, may vary considerably, and do not imply a significant financial gain. When a reviewer does receive a potential perk, it is not in exchange for writing one specific review but rather after a higher-level activity. Lastly, a formerly popular perk was free Google Drive space for reviewers who obtained level 4 in the reviewer hierarchy. Google discontinued this perk prior to the change we investigated. An additional in-depth analysis of other rebates and perks shows that the offers were available only for reviewers in certain locations, such as Southeast Asia, India, and big cities in the US. Because we concentrated on sites that were widely spread across Europe in our main analysis, it is likely that the majority of our reviewers came from nations where no rebates or perks were offered. For the individual-level analysis, selecting an initial seed of reviewers based on whether they had reviewed typical tourist sites in Las Vegas ensured that reviewers came from different parts of the US and the world.

5.5.3 Falsification Tests, Autocorrelated Standard Errors, and Site-Level Trends

To further demonstrate the robustness of our main results, Tables 8 through 10 report various robustness checks that follow the recommendations in Bertrand et al. (2004) and Angrist and Pischke (2009) for DiD estimations for the number of textual reviews (Table 8), the textual review length (Table 9), and nontextual reviews (Table 10). First, we conducted a placebo test to assess the plausibility of our results. We restricted the sample to the control group and then randomly assigned a placebo treatment indicator to sites. Estimating Equation (1) with a randomly assigned placebo variable should not indicate a significant effect. Column (1) of Tables 8 through 10 reports the results. The effects are insignificant, as expected. In addition, similarly, we generated a placebo event. In particular, we restricted the sample to the pretreatment period and then set a placebo $TIME$ variable. Column (2) of Tables 8 through 10 shows the resulting coefficients, which are insignificant, as expected. Second, to address concerns of autocorrelation, we followed Bertrand et al. (2004) and aggregated the panel into two periods, such that each site was observed once before the switch and once after. Column (3) of Tables 8 through 10 reports the results, which are consistent with our main results in Table 3. Third, following Angrist and Pischke (2009, p. 238, Equation 5.2.7), we estimated the main model with a site-specific time trend included. Column (4) of Tables 8 through 10 displays the estimates, which are qualitatively unchanged.

Table 8: Robustness Checks for Textual Reviews

	ln(NUMTEXREV)				
	(1)	(2)	(3)	(4)	(5)
GOOGLE x TIME	–0.098 (0.107)	0.029 (0.176)	0.654*** (0.144)	0.702*** (0.130)	
GOOGLE x TREND (Pre-Trend)					–0.009 (0.081)
Site-Level FE	✓	✓	✓	✓	✓
Observations	293	296	148	586	296
Adj. R^2	0.883	0.611	0.931	0.922	0.611

Note: Robust standard errors are in parentheses. *p < 0.05; ** p < 0.01; *** p < 0.001.

Table 9: Robustness Checks for Review Length

	ln(REVLENGTH)				
	(1)	(2)	(3)	(4)	(5)
GOOGLE x TIME	0.032 (0.065)	0.014 (0.053)	0.153** (0.055)	0.109 (0.079)	
GOOGLE x TREND (Pre-Trend)					0.011 (0.025)
Site-Level FE	✓	✓	✓	✓	✓
Observations	293	296	148	148	296
Adj. R^2	0.303	0.920	0.972	0.972	0.920

Note: Robust standard errors are in parentheses. *p < 0.05; ** p < 0.01; *** p < 0.001.

Table 10: Robustness Checks for Nontextual Reviews

	ln(NONTEXREV)				
	(1)	(2)	(3)	(4)	(5)
GOOGLE x TIME	–0.098 (0.107)	0.260 (0.207)	0.862*** (0.183)	0.755*** (0.145)	
GOOGLE x TREND (Pre-Trend)					0.092 (0.096)
Site-Level FE	✓	✓	✓	✓	✓
Observations	293	296	148	586	296
Adj. R^2	0.883	0.619	0.921	0.932	0.883

Note: Robust standard errors are in parentheses. *p < 0.05; ** p < 0.01; *** p < 0.001.

5.5.4 Parallel Trend Assumption

Finally, we pursued two approaches to test the parallel trend assumption. First, following the suggestion of Angrist and Pischke (2009), we restricted the data to the pretreatment time and regressed the outcome variables on an interaction of a time trend ($TREND$) and the treatment variable $GOOGLE$. With a common trend assumption violation, the coefficient would be significant. As can be seen in Column (5) of Tables 7 through 9, the coefficients are insignificant.

Second, we implemented a fully flexible DiD model with treatment lags and leads, as expressed in Equation (4).

$$\ln(Y_{it}) = \alpha + \Sigma_j \beta_j (GOOGLE_i \times TIME_t) + \beta_2 GOOGLE_i + \Sigma_t \beta_t TIME_t + \mu_s + \varepsilon_{its} \qquad (4)$$

where Y_{it} represents our three dependent variables as above and $TIME_t$ indicates the month of observation. As depicted in Figures A1 to A3 in the Appendix, there were no statistically significant pretreatment trends. All robustness checks underline the validity of our main analysis.

6 Discussion and Conclusion

In this study, we sought to understand the effects of self-contained rewards on the content contribution of users. Even though past research had investigated financial and peer-based rewards, empirical field evidence on the effectiveness of self-contained rewards had not been explored. Besides the overall effects of self-contained rewards, we particularly focused on the spillover effects of incentives on unincentivized tasks and on incentive effectiveness across users with heterogenous expertise with the focal platform.

Using data from Google Maps and Tripadvisor, respectively, provided a favorable research environment for analyzing these effects. Because incentivizing textual reviews was changed in Google's reward system from fall 2017 to spring 2018, we were able to investigate how the newly introduced and adapted design features have affected content contribution. Our estimates suggest that the adaption of Google's reward system has significantly increased the number of textual reviews (62%) and textual review length (16%). We further identified—contrary to our expectations—a positive and substantial spillover effect to nontextual reviews, which increased by 82% even though they were not a target of the restructuring. Finally, we found that more effortful incentives (writing long reviews) were particularly effective for users with high reviewing expertise, while the positive spillover effects of incentives to tasks that require relatively low effort (giving a rating without a text) were particularly effective for users with low reviewing expertise. Importantly, our findings suggest that the introduction of self-contained rewards neither induces a motivation crowding out among reviewers nor creates a positivity bias for the reviewed locations.

6.1 Contribution to Research

To the best of our knowledge, we are the first to examine the effects of self-contained rewards on the extensive and intensive margin of content contribution. We are also the first to analyze spillovers of incentivized tasks to unincentivized tasks. Although prior literature (Qiao et al., 2020) identified that financial incentives have negative spillover effects on reviewing effort for future, unincentivized activities, we document that self-contained rewards actually can have positive spillover effects on unincentivized tasks. Moreover, we are among the first to consider user heterogeneity in determining the effectiveness of rewards on content contribution. Whereas Sun et al. (2017) investigated how user heterogeneity in terms of social connectedness affects the effectiveness of rewards, we analyzed user heterogeneity in terms of user reviewing expertise.

Comparing our findings to the broad literature of financial rewards (Burtch et al., 2018; Cabral and Li, 2015; Khern-am-nuai et al., 2018; Qiao et al., 2020; Sun et al., 2017; Wang et al., 2016; Yu et al., 2022), we conclude that nonfinancial rewards do not exhibit effort crowding out. Financial rewards have also been found to induce a positivity bias into the numerical rating. Our results suggest that self-contained rewards do not induce statistically significant biases into numerical average ratings. Compared to the literature of peer-based rewards (Burtch et al., 2021; Gallus, 2017; Goes et al., 2016), we found that even self-contained awards can elicit user contributions along the extensive and intensive margin. This suggests that a nonfinancial reward system can also be effective without relying on conditional peer votes.

6.2 Implications for Practice

Our results bear meaningful implications on gamification as a social technology to practice.

First, our results suggest that self-contained rewards represent an effective alternative to financial rewards. Financial rewards come with direct marginal costs, they often crowd out reviewer effort, and they induce a positivity bias into numerical ratings (Khern-am-nuai et al., 2018; Liu and Feng, 2021). Some platforms, such as Yelp and Amazon, even prohibit financial incentives to stimulate user-generated content. The reward system we studied circumvents this shortcoming but still effectively elicits user contributions. As a practical recommendation, platforms may consider replacing financial rewards with self-contained rewards to save costs and reduce bias in the ratings. The latter might be particularly appealing to third-party platforms that emphasize keeping their online reviews unbiased—in contrast to first-party sellers, for instance. This replacement would, however, make targeting of rewards more difficult to the platform. If a platform’s strategy is to target customers of newly launched products with financial rewards to generate reviews, the same strategy may be difficult in self-contained review systems unless certain self-contained rewards are introduced for reviewing new products, for instance.

Second, we raise awareness that platforms might contemplate relaxing peer-based conditions for rewards or making them self-contained without sacrificing the effectiveness of awards in incentivizing effortful user contributions. At the same time, our results inform platforms seeking to incentivize contributions at variety and scale that they can effectively use self-contained reward systems for this purpose.

Third, we inform platforms that the effectiveness of reward schemes depends on the heterogeneity of the user population, as low-expertise users respond to rewards differently than high-expertise users. The reward scheme design should therefore be closely tied to the strategic goals of the platform. Two examples of this could be incentivizing content by new (low-expertise) users or reactivating experienced (high-expertise) users. If low-expertise users shall be targeted, our results suggest that emphasis should be placed on rewarding tasks that do not require too much effort. If experienced users shall be targeted, reward systems can effectively incentivize low as well as high-effort tasks.

Fourth, our results raise awareness for spillovers to unincentivized tasks; therefore, evaluating the success of a reward scheme solely based on the incentivized tasks might fall short in capturing the full effectiveness of the scheme. Without our results, platforms might overlook this aspect. Thus, our results encourage platforms to consider effects on unincentivized tasks when evaluating the effectiveness of reward systems.

Finally, our results showcase that reward schemes should be regularly maintained to keep content contributors engaged with the platform. Simple restructuring provides incentives for content contribution to users, which platforms can leverage. Platforms can also monitor the user expertise of their base to inform whether low- or high-effort tasks should be incentivized. Maintaining and revising the reward scheme design, therefore, should play a vital role in a platform's strategic operations management.

6.3 Limitations and Future Research

Even though the restructuring of the Google Maps reward scheme helps in investigating the effect of self-contained rewards, it also presents a natural limitation regarding the evidence we can glean from it. First, our insights remain silent as to what an optimal reward scheme would look like. Second, restructuring a reward scheme is unlikely a one-time event. However, examining the effect of recurring changes of a reward system on content contribution is beyond the scope of our study and presents a fruitful avenue for future research. Third, Google Maps is an all-purpose platform. Reviewing behavior on more specialized platforms, such as for retail or for dining, might warrant follow-up studies. Fourth, we used a sample of public sites for our analyses because they present a solid foundation for the internal validity of our statistical analysis. Naturally, there are plenty of other locations or businesses that could be analyzed. All of these limitations pave the way for future studies.

References

Accenture 2017. Seeing Beyond the Loyalty Illusion: It's Time you Invest More Wisely. https://www.accenture.com/_acnmedia/pdf-43/accenture-strategy-gcpr-customer-loyalty.pdf. Last visited: April 1, 2022.

Angrist, J. D., and Pischke, J. S. 2009. Instrumental Variables in Action: Sometimes You Get What You Need. *Mostly Harmless Econometrics: An Empiricist's Companion*, 113–220.

APKMirror 2020. Google Maps—Navigate & Explore 10.6.2. https://www.apkmirror.com/apk/google-inc/maps/maps-10-6-2-release/. Last visited: February 9, 2022.

Babić Rosario, A., Sotgiu, F., Valck, K., and de Bijmolt, T. 2016. The Effect of Electronic Word of Mouth on Sales: A Meta-Analytic Review of Platform, Product, and Metric Factors. *Journal of Marketing Research*, 53(3), 297–318.

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), 99–126.

Baumeister, R. F., and Leary, M. R. 1995. The Need to Belong: Desire for Interpersonal Attachments as a Fundamental Human Motivation. *Psychological Bulletin*, 117(3), 497–529.

Bertrand, M., Duflo, E., and Mullainathan, S. 2004. How Much Should We Trust Differences-in-Differences Estimates? *Quarterly Journal of Economics*, 119(1), 249–275.

Burtch, G., He, Q., Hong, Y., and Lee, D. 2021. How Do Peer Awards Motivate Creative Content? Experimental Evidence from Reddit. *Management Science* (forthcoming).

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

Cabral, L., and Li, L. 2015. A Dollar for Your Thoughts: Feedback-Conditional Rebates on eBay. *Management Science*, 61(9), 2052–2063.

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), 511–521.

Cheong, C., Cheong, F., and Filippou, J. 2013. Quick Quiz: A Gamified Approach for Enhancing Learning. In *Proceedings of the Pacific Asia Conference on Information Systems (PACIS)*.

Deci, E. L., and Ryan, R. M. 1985. Cognitive Evaluation Theory. In *Intrinsic Motivation and Self-Determination in Human Behavior*. Boston. MA. USA: Springer. 43–85.

DeMatos, C. A., and Rossi, C. A. V. 2008. Word-of-Mouth Communications in Marketing: A Meta-Analytic Review of the Antecedents and Moderators. *Journal of the Academy of Marketing Science*, 36(4), 578–596.

Dreze, X., & Nunes, J. C. 2009. Feeling Superior: The Impact of Loyalty Program Structure on Consumers' Perceptions of Status. *Journal of Consumer Research*, 35(6), 890–905.

Festré, A., and Garrouste, P. 2014. Theory and Evidence in Psychology and Economics About Motivation Crowding Out: A Possible Convergence? *Journal of Economic Surveys*, 29(2), 339–356.

Filippas, A., Horton, J. J., and Golden, J. (2022). Reputation Inflation. *Marketing Science*, (forthcoming).

Floyd, K., Freling, R., Alhoqail, S., Cho, H. Y., and Freling, T. 2014. How Online Product Reviews Affect Retail Sales: A Meta-Analysis. *Journal of Retailing*, 90(2), 217–232.

FTC 2017. The FTC's Endorsement Guides: What People are Asking. https://www.ftc.gov/business-guidance/resources/ftcs-endorsement-guides-what-people-are-asking. Last visited: April 1, 2022.

Gallus, J. 2017. Fostering Public Good Contributions with Symbolic Awards: A Large-Scale Natural Field Experiment at Wikipedia. *Management Science*, 63(12), 3999–4015.

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

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

Gutt, D., Neumann, J., Zimmermann, S., Kundisch, D., and Chen, J. 2019. Design of Review Systems—A Strategic Instrument to Shape Online Reviewing Behavior and Economic Outcomes. *Journal of Strategic Information Systems*, 28, 104–117.

Hamari, J., Koivisto, J., and Sarsa, H. 2014. Does Gamification Work? A Literature Review of Empirical Studies on Gamification. In *Proceedings of the 47th Hawaii International Conference on System Sciences*. Hawaii, USA.

Heath, C., Larrick, R. P., and Wu, G. 1999. Goals as Reference Points. *Cognitive Psychology*, 38(1), 79–109.

iPa4Fun 2020. Google Maps Version History. https://www.ipa4fun.com/history/39/. Last visited: February 9, 2022.

Kankanhalli, A., Taher, M., Cavusoglu, H., and Kim, S. H. S. 2012. Gamification: A New Paradigm for Online User Engagement. In *Proceedings of the 33rd International Conference on Information Systems*. Orlando, FL. USA.

Khern-am-nuai, W., Kannan, K., and Ghasemkhani, H. 2018. Extrinsic Versus Intrinsic Rewards for Contributing Reviews in an Online Platform. *Information Systems Research*, 29(4), 871–892.

Khern-am-nuai, W., Ghasemkhani, H., Qiao, D., and Kannan, K. N. 2022. The Impact of Online Q&As on Product Sales: The Case of Amazon Answer. *SSRN Working Paper*. Available at https://ssrn.com/abstract=2794149. Last visited: March 8, 2022.

King, R., Racherla, P., and Bush, V. 2014. What We Know and Don't Know About Online Word-of-Mouth: A Review and Synthesis of the Literature. *Journal of Interactive Marketing*, 28(3), 167–183.

Kuang, L., Huang, N., Hong, Y., and Yan, Z. 2019. Spillover Effects of Financial Incentives on Non-Incentivized User Engagement: Evidence from an Online Knowledge Exchange Platform. *Journal of Management Information Systems*, 36(1), 289–320.

Lazarus, R. S. 1991. Cognition and Motivation in Emotion. *American Psychologist*, 46(4), 352.

Liu, D., Santhanam, R., and Webster, J. 2017. Toward Meaningful Engagement: A Framework for Design and Research of Gamified Information Systems. *MIS Quarterly*, 41(4). 1011–1034.

Liu, Y., and Feng, J. 2021. Does Money Talk? The Impact of Monetary Incentives on User-Generated Content Contributions. *Information Systems Research*, 32(2), 394–409.

Locke, E. A., and Latham, G. P. 1990. *A Theory of Goal Setting & Task Performance*. Prentice-Hall, Inc.

Lou, J., Fang, Y., Lim, K. H., and Peng, J. Z. 2013. Contributing High Quantity and Quality Knowledge to Online Q&A Communities. *Journal of the American Society for Information Science and Technology*, 64(2), 356–371.

Moro S., Ramos, P., Esmerado, J., and Jalali, S. M. J. 2019. Can We Trace Back Hotel Online Reviews' Characteristics Using Gamification Features? *International Journal of Information Management*, 44, 88–95.

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

Przybylski, A. K., Rigby, C. S., and Ryan, R. M. 2010. A Motivational Model of Video Game Engagement. *Review of General Psychology*, 14(2), 154–166.

Qiao, D., Lee, S.-Y., Whinston, A. B., and Wei, Q. 2020. Financial Incentives Dampen Altruism in Online Prosocial Contributions: A Study of Online Reviews. *Information Systems Research*, 31(4), 1361–1375.

Racherla, P., and Friske, W. 2012. Perceived "Usefulness" of Online Consumer Reviews: An Exploratory Investigation Across Three Services Categories. *Electronic Commerce Research and Applications*, 11, 548–559.

Reimer, T., and Benkenstein, M. 2016. Altruistic eWOM Marketing: More Than an Alternative to Monetary Incentives. *Journal of Retailing and Consumer Services*, 31, 323–333.

Reviewtrackers. 2021. *Online Reviews Statistics and Trends*. https://www.reviewtrackers.com/reports/online-reviews-survey/. Last visited: February 9, 2022.

Ryan, R. M., and Deci, E. L. 2000. Self-Determination Theory and the Facilitation of Intrinsic Motivation, Social Development, and Well-Being. *American Psychologist*, 55(1), 68–78.

Ryan, R. M., and Deci, E. L. 2002. Overview of Self-Determination Theory: An Organismic Dialectical Perspective. In *Handbook of Self-Determination Research*, E. L. Deci and R. M. Ryan (eds.). Rochester, NY. USA: Rochester University Press, 3–33.

Santhanam, R., Liu, D., and Shen, W. C. M. 2016. Research Note—Gamification of Technology-Mediated Training: Not All Competitions Are the Same. *Information Systems Research*, 27(2), 453–465.

Schöbel, S., Janson, A., Jahn, K., Kordyaka, B., Turetken, O., Djafarova, N., Saqr, M., Wu, D., Söllner, M., Adam, M., Gad, P. H., Wesseloh, H., and Leimeister, J. M. 2020. A Research Agenda for the Why, What, and How of Gamification Designs: Outcomes of an ECIS 2019 Panel. *Communications of the Association for Information Systems*, 46, 706–721.

Smith, M. 2002. Evaluation, Uncertainty and Motivation. *Ethical Theory and Moral Practice*, 5(3), 305–320.

Stephen, A., Bart, Y., Du Plessis, C., and Goncalves, D. 2012. Does Paying for Online Product Reviews Pay Off? The Effects of Monetary Incentives on Content Creators and Consumers. In *NA–Advances in Consumer Research*, 40, 228–231.

Stock, R. M., Oliveira, P., and von Hippel, E. 2015. Impacts of Hedonic and Utilitarian User Motives on the Innovativeness of User-Developed Solutions. *Journal of Product Innovation Management*, 32(3), 389–403.

Sun, Y., Dong, X., and McIntyre, S. 2017. Motivation of User-Generated Content: Social Connectedness Moderates the Effects of Monetary Rewards. *Marketing Science*, 36(3), 327–470.

von Rechenberg, T., Gutt, D., and Kundisch, D. 2016. Goals as Reference Points: Empirical Evidence from a Virtual Reward System. *Decision Analysis*, 13(2), 153–171.

Wang, J., Ghose, A., and Ipeirotis, P. 2012. Bonus, Disclosure, and Choice: What Motivates the Creation of High-Quality Paid Reviews. In *Proceedings of the 33rd International Conference on Information Systems*. Orlando, USA.

Wang, J., Khoo, A., Liu, W. C., and Divaharan, S. 2008. Passion and Intrinsic Motivation in Digital Gaming. *CyberPsychology & Behavior*, 11(1), 39–45.

Wang, S., Pavlou, P. A., and Gong, J. 2016. On Monetary Incentives, Online Product Reviews, and Sales. In *Proceedings of the 37th International Conference on Information Systems*. Dublin, Ireland.

Wang, X., and Sanders, G. L. 2019. For Money, and for Fun: Exploring the Effects of Gamification and Financial Incentives on Motivating Online Review Generation. In *Proceedings of the 25th Americas Conference on Information Systems*. Cancún, Mexico.

Wang, Y., Goes, P., Wei, Z., and Zeng, D. 2019. Production of Online Word-of-Mouth: Peer Effects and the Moderation of User Characteristics. *Production and Operations Management*, 28(7), 1621–1640.

Wang, L., Gunasti, K., Shankar, R., Pancras, J., and Gopal, R. 2020. Impact of Gamification on Perceptions of Word-of-Mouth Contributors and Actions of Word-of-Mouth Consumers. *MIS Quarterly*, 44(4), 1987–2011.

Yu, Y., Khern-am-nuai, W., and Pinsonneault, A. 2022. When Paying for Reviews Pays Off: The Case of Performance-Contingent Monetary Rewards. *MIS Quarterly*, 46(1), 609–626.

Appendix

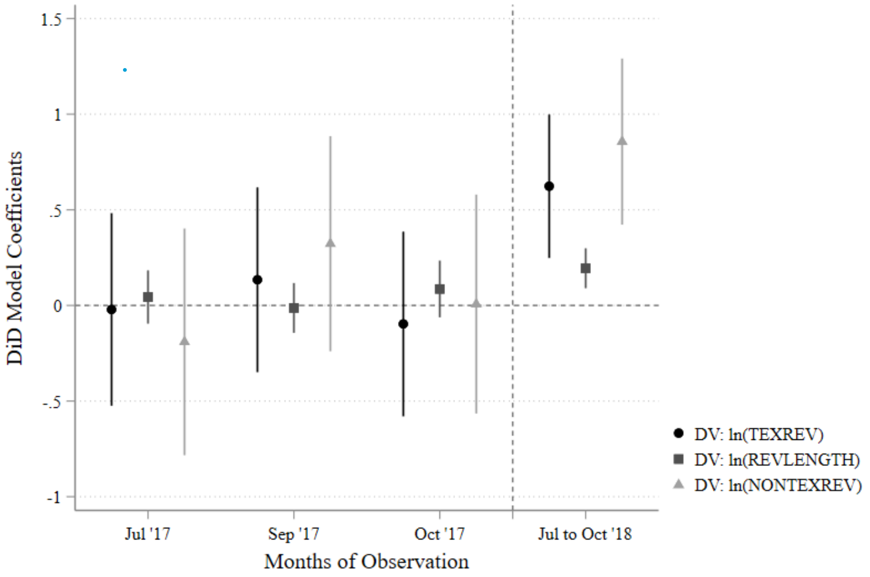

Figure A1. Pre-Trend Analysis with Treatment Leads Full Sample

Note: This figure plots the coefficient of the interactions of the months before the restructuring and *GOOGLE*. The months after restructuring have been binned together. August '17 serves as the omitted baseline.

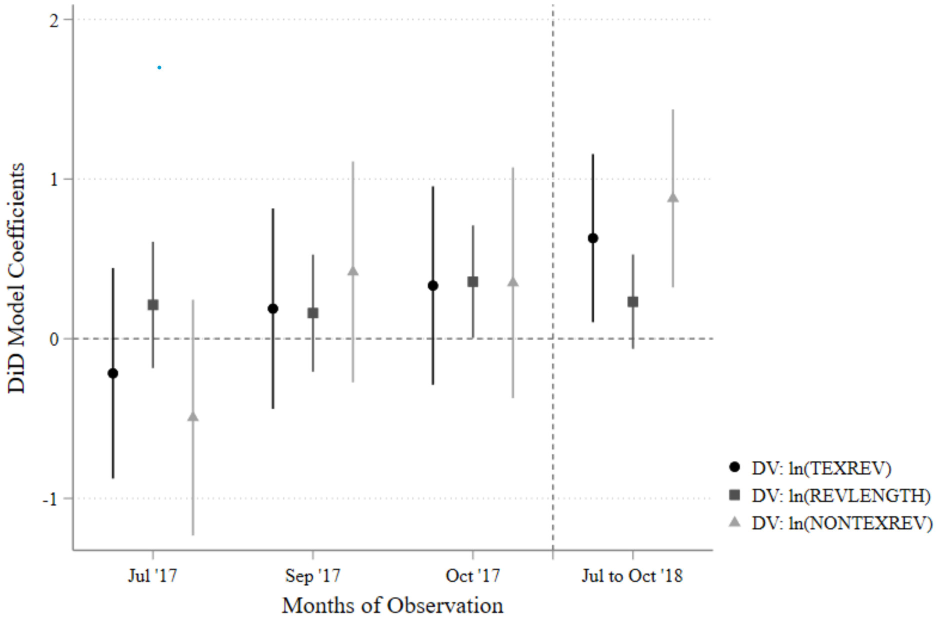

Figure A2. Pre-Trend Analysis with Treatment Leads for Users with Low Reviewing Expertise

Note: This figure plots the coefficient of the interactions of the months before the restructuring and *GOOGLE*. The months after restructuring have been binned together. August '17 serves as the omitted baseline.

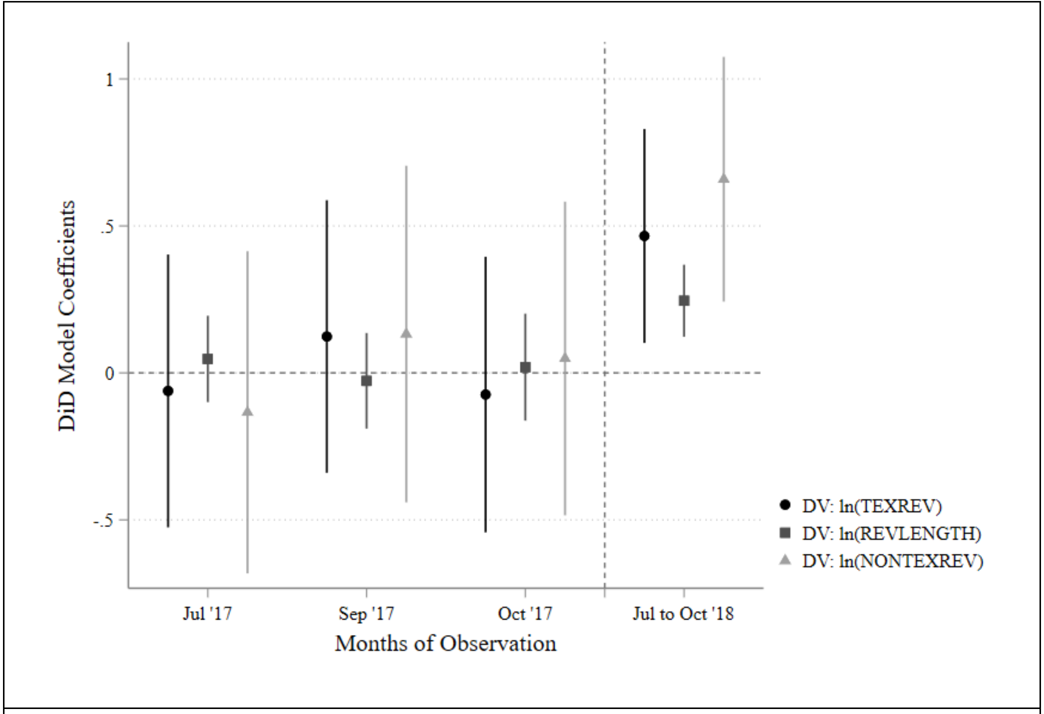

Figure A3. Pre-Trend Analysis with Treatment Leads for Users with High Reviewing Expertise

Note: This figure plots the coefficient of the interactions of the months before the restructuring and *GOOGLE*. The months after restructuring have been binned together. August '17 serves as the omitted baseline.