MIS Quarterly

RESEARCH NOTE

REVIEW MODERATION TRANSPARENCY AND ONLINE REVIEWS: EVIDENCE FROM A NATURAL EXPERIMENT[1]

Lianlian (Dorothy) Jiang
Department of Decision & Information Sciences, Bauer College of Business, University of Houston,
Houston, TX, U.S.A. {ljiang@central.uh.edu}

T. Ravichandran
Lally School of Management & Technology, Rensselaer Polytechnic Institute (RPI),
Troy, NY, U.S.A. {ravit@rpi.edu}

Jason Kuruzovich
Lally School of Management & Technology, Rensselaer Polytechnic Institute (RPI),
Troy, NY, {kuruzj@rpi.edu}

This paper empirically investigates how review moderation transparency affects the volume, length, and negativity of reviews. A change to the Yelp platform in 2010, introducing review moderation and displaying filtered reviews, created a natural experiment. We used a panel dataset of online reviews from the same set of restaurants on both the Yelp and TripAdvisor platforms in a difference-in-differences (DID) model to test how review moderation transparency affected our outcome variables. We found that increasing review moderation transparency negatively affects review volume but positively affects review negativity. The results also indicate that providing review moderation transparency reduces review length, especially for reviews with positive sentiment. Our findings suggest that providing review moderation transparency induces users to invest less effort in review contributions, especially when they are submitting positive reviews. We discuss the theoretical and practical implications of these results as they relate to the design and use of online review platforms.

Keywords: Review moderation transparency, online reviews, reduction in contribution investment, natural experiment, difference-in-differences

Introduction

Review platforms such as Google, Facebook, Yelp, and TripAdvisor have successfully leveraged online reviews, integrating them with nearly all facets of digital life—from self-expression (Gosling et al., 2011) to purchasing decisions (Moe & Schweidel, 2012). The use of platforms by Russians for election tampering (Shane, 2018) and the removal of political leaders from platforms (Klepper & O'Brien, 2021) highlight both the power held by platforms and their responsibility to be transparent about moderation processes. Review platforms are also under increased pressure from the retailers that rely upon them to develop and enforce review content guidelines that prevent fraudulent content and maintain content quality (Vaughan, 2020). Review filtering on platforms such as Yelp has become very important, and the need to maintain review quality through moderation is recognized by all platform stakeholders (Luca & Zervas, 2016).

However, the majority of review platforms need to be more transparent about whether and how they moderate reviews. This has led to heightened concern among consumers that platform operators could manipulate reviews to increase profits under the pretext of filtering out fraudulent content (Erskine, 2017). For example, Yelp has received over 2000

[1] H. Raghav Rao was the accepting senior editor for this paper. Jui Ramaprasad served as the associate editor.

DOI:10.25300/MISQ/2023/16216

FTC complaints that review filtering is handled differently for advertisers and non-advertisers (Snyder, 2015), and the European Commission has summarized a variety of issues related to platform transparency, suggesting the potential for future regulation (European Commission, 2018). As an initial step, in 2018, a partnership between 12 major companies—including Apple, Facebook (Meta), Google, Reddit, and Twitter—and activists endorsed the Santa Clara Principles to provide content moderation transparency (York, 2020). However, it is unclear how the awareness of moderation will impact user behavior, as some platforms, such as TripAdvisor and Amazon, have a policy of not offering any specifics on review moderation. Other platforms, such as Yelp, are transparent about their review moderation, enabling users to see filtered reviews and providing videos explaining their review moderation policies (Yelp, 2010).

Given the importance of review contribution to review platform success, there is a critical need for additional research on review moderation transparency and its potential effects. Providing more information about review moderation may remind users of the scrutiny and censorship imposed by platforms' filtering systems, increasing users' concerns that their reviews may be filtered out and potentially reducing their investment of time and effort in contributing reviews. Alternatively, disclosing more information about review moderation may incentivize users to invest more care and effort in creating reviews in order to make them less likely to be removed by the filtering system. Further, increased transparency in the review moderation process could incentivize changes in the behaviors of individuals submitting fake reviews. Given these different potential consequences of review moderation transparency on reviewing behaviors, in this paper, we investigate how review moderation transparency affects review volume (Huang et al., 2017), length (Mudambi & Schuff, 2010), and negativity (Huang et al., 2017). We use *review moderation transparency* in the remainder of this paper to refer to the disclosure of information regarding why and how reviews are filtered and whether reviews are filtered consistently. As review characteristics are critical to the value created by review platforms, this suggests that transparency may be a vital design decision of online platforms.

Our study used a panel dataset of 1,016 of the same restaurants' reviews from two leading online review platforms (Yelp and TripAdvisor) from 2008-2012 to empirically test the impact of review moderation transparency on review platforms. A substantial change to the Yelp platform in 2010, which provided information about its review filtering system's principles, created a natural experiment. We used a difference-in-differences (DID) approach to study the differential effect of providing review moderation transparency on Yelp versus a control group (TripAdvisor, which did not change its platform during our observational period and is not transparent about how it filters reviews). Employing identical restaurant reviews on two separate venues in our analysis enabled us to control for unobserved restaurant differences between the treatment and control groups.

This research makes several contributions. First, this study is one of the first to provide empirical evidence of the relationship between the information transparency strategy of an online platform and users' content contribution behaviors. Second, we provide robust empirical evidence that review moderation transparency has a complex impact on review quality by enhancing review negativity and reducing review length, as numerous previous studies have found that review negativity and length positively affect review quality (Cao et al., 2011; Chen & Lurie, 2013; Mudambi & Schuff, 2010).[2] In addition, we found that review moderation transparency reduces review volume, which, along with review length, is an indicator of overall user involvement on the platform. Our paper thus sheds light on the nuanced effects of transparency as a design choice of review platforms that can impact users' reviewing behaviors. By examining review moderation transparency in a robust empirical context involving a natural experiment, we contribute to the literatures on information transparency and online review platforms and inform practice by highlighting the importance of moderation transparency as a design parameter of online platforms.

Related Literature

We draw on the literature on online review platforms and information transparency as a foundation to identify gaps in the literature and frame our study's contribution.

Online Review Platforms

Many aspects of online review platforms have been extensively studied. A number of studies have examined the effects of online reviews on product sales (e.g., Chevalier & Mayzlin, 2006; Duan et al., 2008) and brand perceptions (Sparks & Browning, 2011). A related stream has also examined the characteristics of online reviews (e.g., Kuan et

[2] Since the relationship between review length, review negativity, and review helpfulness has been demonstrated numerous times, review helpfulness is not a dependent variable in our study.

al., 2015; Mudambi & Schuff, 2010) and reviewers (e.g., Ghose & Ipeirotis, 2011) that impact review helpfulness. Another stream has examined review generation or the process of contributing reviews, with researchers investigating the impact of prior reviews (e.g., Chen et al., 2018; Moe & Schweidel, 2012), users' selection bias (e.g., Hu et al., 2006; Hu et al., 2009), and the social influence of online friends or family members enabled by social network integration (e.g., Huang et al., 2017). While past studies have found that other users impact users' reviewing behaviors (Goes et al., 2014; Huang et al., 2017), with online platforms increasingly engaging in review moderation, there remains a need to understand the effects of scrutiny by non-socially proximal agents on users' information sharing behaviors, which is our paper's focus.

Information Transparency and Review Moderation Transparency

Information transparency is a crucial subject studied widely in IS (Schnackenberg & Tomlinson, 2016). Despite the breadth and depth of transparency literature, there are critical gaps in the literature, especially in the rapidly evolving area of online platforms. First, prior studies have primarily explored the positive outcomes of adopting a transparent information strategy, implying that more transparency is nearly always beneficial for a firm (e.g., Trifts & Häubl, 2003; Wang & Benbasat, 2016) and using either stylized analytical models or experiments in simplified settings to derive findings. It is possible that in the context of online platforms, information transparency may have different implications when viewed from the perspectives of the operator, user, or platform contributor. As a result, to develop sound transparency strategies, it is essential to take a more nuanced view and use observational data to study information sharing in actual firm-customer interactions.

Second, prior research has largely focused on how information transparency strategy affects purchases (e.g., Oh & Lucas Jr, 2006; Shulman et al., 2015) where information transparency is used as a means to make users more informed to generate trust and thereby enhance purchase intentions. In contrast, our study context is platforms where reviews are hosted by an intermediary that generates revenue indirectly through ads (e.g., Yelp) and where users primarily consume the information presented or contribute information to the platform. Hence, the findings from studies on electronic commerce contexts may not generalize to information intermediary platforms; thus, there is a need to better understand the transparency strategy of such platforms.

Filtering reviews is common for review platforms, but not all users are aware of the practice (Patterson, 2016). In order to ensure that reviews comply with an online platform's posting guidelines and to protect users from fake, unhelpful, or malicious content, platforms primarily rely on automated sophisticated software referred to as a *review filtering system* to flag and filter reviews (Luca & Zervas, 2016). Once a review is flagged, it will be filtered out, will not be displayed on a business page, and will not be factored into the overall rating for a business.

However, only transparent platforms provide information about the principles and the logic underlying the filtering system and flagged reviews. In contrast, a platform with no review moderation transparency, such as TripAdvisor, may not display removed reviews nor explain how its review filter works. We conceptualize *review moderation transparency* as the extent to which a review platform discloses information about *how* a review filtering system works, *why* reviews are filtered out, and whether different kinds of reviews are screened *consistently* (Wang & Benbasat, 2007). The how and why of review moderation can be implemented by explaining the general filtering process, describing the characteristics of the reviews that are likely to be filtered out, and emphasizing the filtering policy's benefit to users. Transparent platforms also generally provide information for users (e.g., enabling users to see the blocked reviews that the filtering system has removed) to verify the consistency of its filtering system.

Content moderation has been studied in other contexts. The principles on which moderation is based have been studied using user comments from newspaper forums (Boberg et al., 2018). Users' desire for more transparency in content moderation after their comments have been moderated has been demonstrated via surveys (Suzor et al., 2019). The investigation of flagged content on Reddit reveals a lack of transparency in moderation practices, as Reddit does not provide any notification or explanation for removing content (Juneja et al., 2020). Users often feel powerless regarding content moderation, and user-moderated platforms are typically perceived as suffering from a lack of transparency equivalent to commercially moderated platforms (Cook et al., 2021).

Closely related work by Jhaver et al. (2019) examined how the specific explanation for removing a Reddit post affects the user's future posting likelihood and future rates of having posts removed. To the best of our knowledge, users' posting behaviors in response to the provision of review moderation transparency at a *platform level* in terms of aggregate user behaviors have not been studied. First, our study is distinct because it investigates the resulting aggregate changes in user behavior across the online review platform resulting from the transparency of the content moderation *policy*, rather than changes resulting from content moderation *actions* involving users whose reviews were filtered out. Second, our outcome variables (including review volume, length, ratings, and sentiment) also differ from those of Jhaver et al. (2019), which looked at user-level outcomes (including future submission and future removal). Third, the percentage of users impacted by content moderation actions conducted by the platform is likely to be very small. In contrast, an average user may become aware of the content moderation policy after the platform provides more information about the policy. Therefore, it is an open question whether users will be affected by changes to the content moderation policy. Understanding this at a deeper level has significant implications for content moderation policy and, consequently, for effective platform design. Fourth, we provide evidence for the underlying mechanism influencing our outcome variables, making this paper distinct from prior work.

The Impact of Review Moderation Transparency

When users are unaware that review platforms moderate reviews, disclosing relevant information about the filtering policy will confirm its existence. By providing information about the consistency with which positive and negative reviews or reviews from advertisers and non-advertisers are filtered, a platform can signal that the filtering system will evaluate all reviews similarly. This is likely to reinforce the salience of the scrutiny of all reviews, irrespective of their nature, by the review filtering system and raise awareness that the filtering system may remove review submissions (Prentice-Dunn & Rogers, 1982).

On the one hand, this increased scrutiny may reduce and even minimize the effort users invest in contributing reviews. According to expected utility theory (Rabin, 2013), if users are concerned that their reviews will not get through the filtering system, they may invest less time and effort in review creation in order to avoid significant loss of effort, such as submitting reviews that are subsequently flagged and removed. This concern could lead to shorter reviews (reducing average review length). Further, the majority of fake reviews from businesses are typically overly positive (Luca & Zervas, 2016). As a result, users may believe the system is more likely to remove positive reviews than negative ones. Thus, users may contribute less content in positive reviews than in negative reviews (reducing review length). This awareness of filtering may lead to a more significant reduction of positive (vs. negative) reviews, as in some cases, users may even decide to forego contributing to the platform (reducing overall review volume), particularly in the case of positive reviews. Thus, users' perceptions about filtering patterns could induce a greater reduction of positive reviews, which could decrease the average sentiment of reviews and online ratings. It should be noted that filtering systems typically do not explicitly target positive reviews. Instead, positive reviews may be filtered out at a higher rate because fake reviews tend to be predominantly positive. We refer to this potential mechanism of the impact of review moderation transparency on user behaviors as a *reduction in contribution investment.*

On the other hand, the increased scrutiny of providing review moderation transparency could induce users to increase their effort in contributing to the platform to ensure that their contribution is not flagged for removal by the system. As the moderation system's goal is to remove fraudulent reviews, users may attempt to increase efforts to incorporate more authentic experiences with products to prove that they are valid reviews that should not be flagged for removal (Solis, 2018) (increasing average review length). Further, if users believe that review platforms' automated filters are more likely to remove positive reviews, users may increase the length of positive reviews to ensure that their reviews will not be flagged for removal. This could even lead to users submitting more positive reviews to ensure that their reviews are published, as they may believe it to be more likely that the filtering system will remove posts from users with a single review, suspecting that it is from a bot or a fake profile (increasing online ratings and average review sentiment) (Solis, 2018; Yelp, 2010). Users submitting more reviews to give their account increased credibility may subsequently increase review volume. We refer to this potential mechanism associated with review moderation transparency as *increased contribution diligence.*

The above discussions pertain to typical expected behaviors in response to users becoming aware of the platform's review moderation. In contrast to regular users, those submitting fake reviews might have very different motivations and incentives that shape their behaviors. For fake/paid reviewers, increased transparency might lead them to revise their strategy to ensure that their contributions will not be filtered out, such as contributing detailed reviews that appear authentically crafted. Alternatively, increased transparency could lead fake reviewers to minimize potential losses in terms of effort by contributing multiple shorter reviews. While each of these mechanisms is possible, we believe that aggregate user behavior is unlikely to be significantly shifted by any slight change in fake reviewer behavior because fake reviewers typically constitute a small proportion of reviewers (likely less than 20%) on review platforms (Luca & Zervas, 2016). Further, platforms are likely to use sophisticated filtering software that relies on user history and context data (such as location) to validate the users' contribution in addition to the text. As a result, while the potential change of fake reviewer behavior resulting from transparency is an interesting avenue for study, we do not directly address this mechanism in the paper.

In summary, providing more information about review moderation practice may increase user awareness of review moderation and impact the pattern of review contribution on the platform. It is unclear whether this shift will result in a decrease or an increase in users' investment in review contributions. In the following section, we discuss our empirical testing of the effects of review moderation transparency on review volume, length, review sentiment, and online ratings, shedding light on the relative dominance of the identified mechanisms.

Research Methodology

Empirical Setting

We tested our hypotheses using data from Yelp and TripAdvisor, two leading consumer review websites with different review moderation transparency strategies. Prior to March 18, 2010, Yelp had review filters in place but did not display the filtered reviews nor discuss the algorithm or logic used to conduct the filtering. On March 18, 2010, Yelp added a video designed to help its users understand why its review filter exists and how it works (Yelp, 2010) (See Figure 2). For example, the video mentions that both positive and negative reviews can be affected. Inappropriate activities that could initiate filtering include owners aggressively soliciting 5-star reviews for their businesses, manufacturing fake 5-star reviews about themselves, or writing 1-star reviews about their competitors. On April 5, 2010, Yelp changed its website by adding a section displaying reviews that have been filtered out. Recommended reviews for Yelp are those that pass their evaluation and are not flagged and filtered out. Yelp's website (shown in Figure 1) separates recommended reviews from reviews that have been filtered out (Yelp labels them non-recommended reviews), with recommended reviews placed just below the brief introduction section for each business. Only reviews that the filtering system has screened out are included in the set of non-recommended reviews. In order to check a flagged review, a user has to scroll to the bottom of a business page and click the associated link, as shown in Figure 2. The link's title indicates how many reviews are currently not recommended and therefore screened out. Recommended and non-recommended reviews are clearly differentiated by color, with recommended reviews' ratings in red and non-recommended reviews' ratings in gray.

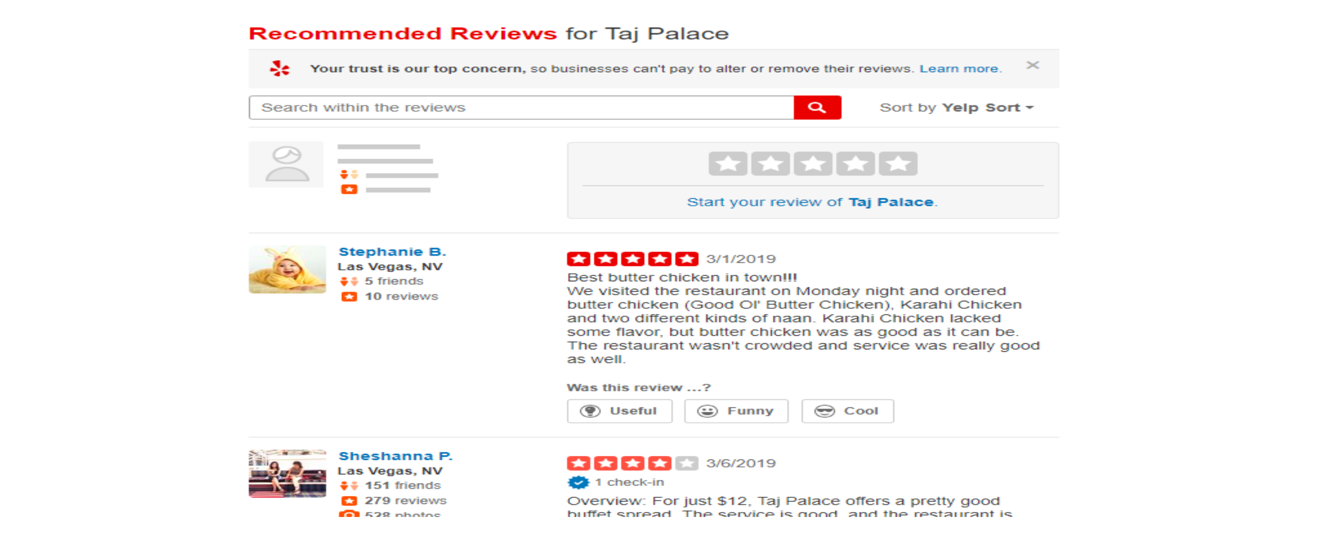
In contrast to Yelp, TripAdvisor's review moderation processes remained unchanged during our observational period. In essence, while both platforms moderate reviews, Yelp made substantial changes to its platform to become very transparent about its policies and deliberately ensure that its users are aware of these policies. In contrast, TripAdvisor is opaque about its moderation policies and logic.

Observational Data and Measurement

We obtained Yelp's reviews from the Yelp dataset, a subset of their user data for personal, educational, and academic purposes. The dataset includes 10 U.S. cities, and we selected six of the largest cities: Las Vegas, Philadelphia, Phoenix, Charlotte, Cleveland, and Tempe. We did not select the remaining four cities because they are located very close to and overlap with the selected cities. We then programmatically collected the same set of restaurant reviews from TripAdvisor. The reviews we used for our study are posted reviews. Our sample does not include reviews that have been filtered out, as TripAdvisor does not disclose these reviews. All of the restaurants in these six cities with accounts on both platforms and reviews posted before and after the system change at Yelp were included. We pooled the reviews for the set of restaurants that appeared on both platforms to create our treatment (Yelp) and control (TripAdvisor) groups.

We measured review volume as the monthly total number of reviews posted on a given restaurant's profile page and review length by the number of words a review contained. Review negativity was measured by negative sentiment in reviews and online ratings. We measured sentiment using the Microsoft Azure Text Analytics application programming interface (API). The Microsoft Text Analytics API evaluates text input by using a machine learning algorithm to generate a numeric sentiment score ranging from 0 (negative) to 1 (positive). To measure negative sentiment in reviews more directly, we recorded it to range from 0 (positive) to 1 (negative). We used the star rating to measure online ratings, a numeric score ranging from 1 (negative) to 5 (positive) for each review.

Our panel was composed of monthly data at the restaurant level and we removed the months with no reviews posted. To aggregate data at the monthly level, we first measured the review length and review negativity of each review and then averaged them across reviews in each month for each restaurant. To calculate review volume, we similarly summed each restaurant's total number of reviews each month. The descriptive statistics and a correlation matrix of the main variables are presented in Tables 1 and 2.



Recommended Reviews for Taj Palace

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

Search within the reviews

Sort by **Yelp Sort**

Start your review of **Taj Palace**.

Stephanie B.
Las Vegas, NV
5 friends
10 reviews

3/1/2019

Best butter chicken in town!!!
We visited the restaurant on Monday night and ordered butter chicken (Good Ol' Butter Chicken), Karahi Chicken and two different kinds of naan. Karahi Chicken lacked some flavor, but butter chicken was as good as it can be. The restaurant wasn't crowded and service was really good as well.

Was this review ...?

Useful Funny Cool

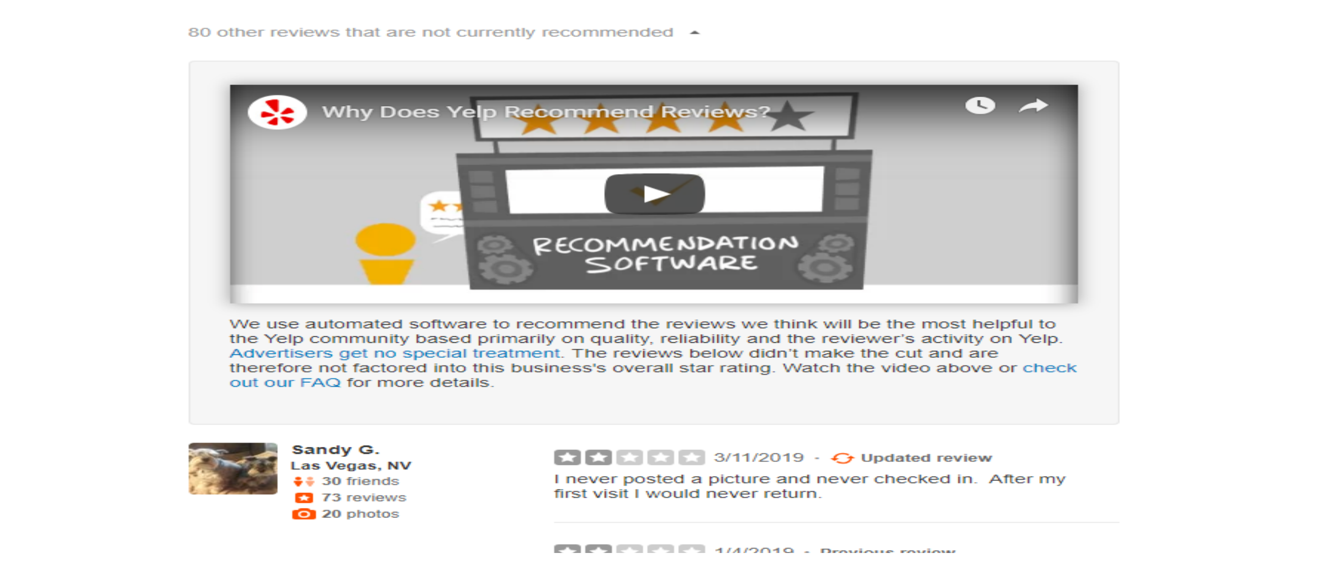
Sheshanna P.
Las Vegas, NV
151 friends
279 reviews
528 photos

3/6/2019

1 check-in

Overview: For just $12, Taj Palace offers a pretty good buffet spread. The service is good, and the restaurant is

Figure 1. Web Page with Recommended Reviews on Yelp.com (left)



80 other reviews that are not currently recommended

We use automated software to recommend the reviews we think will be the most helpful to the Yelp community based primarily on quality, reliability and the reviewer's activity on Yelp. Advertisers get no special treatment. The reviews below didn't make the cut and are therefore not factored into this business's overall star rating. Watch the video above or check out our FAQ for more details.

Sandy G.
Las Vegas, NV
30 friends
73 reviews
20 photos

3/11/2019 · Updated review

I never posted a picture and never checked in. After my first visit I would never return.

1/4/2019 · Previous review

Figure 2. Web Page with the Feature of Disclosing Filtered Reviews to Users and the Explanation Video of Review Recommendation Software on Yelp.com (right)

Table 1. Descriptive Statistics

Variables	*N*	Mean	*SD*	Min	Median	Max
Review volume	41176	3.187	4.872	1	2	129
Ln (review length)	41176	4.555	0.837	0	4.684	6.878
Negative sentiment	41176	0.277	0.268	0	0.169	0.999
Online ratings	41176	3.787	0.987	1	4	5
Restaurant age	41176	3.386	1.668	0	3.357	8.064
Prior overall rating	41176	3.797	0.545	1	3.862	5
Prior review volume	41176	57.846	104.323	0.5	26.5	1852.5

Table 2. Correlation Matrix

Variables	Review volume	Review length	Negative sentiment	Online ratings	Restaurant age	Prior rating	Prior review volume
Review volume	1						
Ln (review length)	0.104	1					
Negative sentiment	-0.003	0.107	1				
Online ratings	0.045	-0.120	-0.593	1			
Restaurant age	0.257	0.042	0.011	0.026	1		
Prior overall rating	0.060	-0.032	-0.193	0.357	-0.022	1	
Prior review volume	0.793	0.129	0.015	0.021	0.372	0.034	1

Identification Strategy and Estimation Model

Our identification strategy relies on a natural experiment in which Yelp implemented two consecutive system design changes related to review moderation transparency on March 18, 2010, and April 5, 2010. We treated the two consecutive changes on Yelp as one exogenous shock to platform users. TripAdvisor (our control group) had low levels of review moderation transparency during this period, as the website did not provide any specifics about its review moderation mechanisms. Since March and April 2010 are the policy change months, we excluded these months' data from the analysis. The pretreatment and posttreatment periods are 24 months each. To better control for the effects of omitted variables (e.g., restaurant quality, restaurant tenure, menu change), difference-in-differences (DID) estimation is utilized to estimate the impact of providing review moderation transparency on the volume and characteristics of online reviews for the same set of restaurants.

We estimated DID models to evaluate the treatment effect of the system change from no review moderation transparency to providing review moderation transparency for the same set of restaurants, reflected by Equation (1). We used ordinary least squares (OLS) regression to test Equation (1). Since review length (which ranges from 1 to 919) is skewed to the right and has a comprehensive data range over several orders of magnitude, we log-transformed the value of review length.

$$y_{itp} = \beta_1 Post_t \times Treat_p + X_{itp} + \alpha_{ip} + \gamma_t + \varepsilon_{\ itp} \qquad (1)$$

In the above Equation (1), *i* indicates the restaurant, *t* denotes the month, and *p* indicates a restaurant is listed on Yelp or TripAdvisor. The key parameter of interest in these models, beta sub 1, captures the impact of review moderation transparency. *Treat* is a dummy variable to indicate the treatment group (i.e., Yelp). *Post* is a dummy variable set to one if the reviews are made after the platform change and otherwise set to zero. The dependent variables in Equation (1) are the t^{th} month review volume, review length, online ratings, and review negativity of the reviews submitted for restaurant *i* listed on platform p. We added a restaurant fixed effect α_{ip} and a year-month time fixed effect γ_t. ε_{itp} indicates the error term.

Because the fixed effects absorb any time-invariant control variable, we followed prior studies (Chen et al., 2018; Huang et al., 2017) in this stream and added three time-variant control variables in Equation (1), represented by X_{itp}: restaurant age represents how many years a restaurant has had an account on Yelp, prior overall rating of a restaurant represents the overall rating of a restaurant before a review is posted, and prior review volume represents the number of reviews posted to a restaurant before a review is submitted. We controlled for *prior overall rating of a restaurant* because many studies have found that *prior overall rating* of a product influences a user's rating. *Prior review volume* was controlled for because more reviews may enhance users' impressions of the restaurant, which may contribute to adjusting a user's rating. In addition, if a restaurant has had an account on a review platform for many years, it may influence a user's impression of the restaurant and consequently affect their reviewing behaviors.

Results

Main Findings

Our results on the effect of review moderation transparency are presented in Table 3. We detected a negative and significant impact of review moderation transparency on review volume (β = -1.786, $p < 0.001$). Column 2 of Table 3's result indicates that providing moderation transparency decreased review length (β = -0.416, $p < 0.001$). Column 3 of Table 3 suggests that providing review moderation transparency increased negative sentiment in reviews (β = 0.012, $p < 0.1$). Review moderation transparency is negatively related to online ratings (β = -0.184, $p < 0.01$) in Column 4 of Table 3. These findings suggest that the *reduction in contribution investment* is more likely to be the dominant mechanism.

Table 3. Effect of Review Moderation Transparency on Review Volume and Review Characteristics

Variables	(1) Review volume	(2) Ln (review length)	(3) negative sentiment	(4) Online ratings
Post × Treatment	-1.786***(0.213)	-0.416***(0.026)	0.012+(0.007)	-0.184***(0.026)
Observations	41176	41176	41176	41176
R-squared	0.697	0.309	0.134	0.24
Number of restaurants	1016	1016	1016	1016
Restaurant fixed effect	Yes	Yes	Yes	Yes
Time fixed effect	Yes	Yes	Yes	Yes
Restaurant controls in Table 1	Yes	Yes	Yes	Yes

Note: Robust Standard errors in parentheses (clustered on restaurants) ****p* < 0.001, ***p* < 0.01, **p* < 0.5, +p<0.1 for Tables 3-8. "Restaurant controls in Table 1" indicates that we controlled for restaurant age, prior review volume, and average prior overall ratings of restaurants.

Table 4. Effect of Review Moderation Transparency on Review Characteristics

Variables	(1) Volume of negative ratings	(2) Volume of positive ratings	(3) Length of review with negative sentiment	(4) Length of review with positive sentiment	(5) Online ratings
High_overall × Post × Treatment					-0.202***(0.050)
Post × Treatment	0.048+(0.025)	-1.615***(0.195)	-0.234***(0.037)	-0.444***(0.028)	-0.085*(0.040)
High_overall × Post					-0.079+(0.042)
Observations	41176.000	41176	17620	35842	41176
R-squared	0.411	0.649	0.277	0.31	0.243
Number of restaurants	1016	1016	1016	1016	1016
Restaurant fixed effect	Yes	Yes	Yes	Yes	Yes
Time fixed effect	Yes	Yes	Yes	Yes	Yes
Restaurant controls in Table 1	Yes	Yes	Yes	Yes	Yes

Additional Analyses

Though our main results suggest that a *reduction in contribution investment* is the main mechanism influencing review volume and length, we now discuss several tests related to the relative effect of the treatment on positive and negative reviews, measured by rating, sentiment, and prior history. After filtered reviews are revealed to users as a practice of providing review moderation transparency, users may observe that most reviews that are filtered out on Yelp are positive (Luca & Zervas, 2016). Again, while review filters have no preference for positive reviews, a substantial portion of the pool of fraudulent reviews is composed of overly positive reviews submitted by agents of the business owner. Therefore, users may have concerns when attempting to contribute positive reviews. If we were to observe a greater effect of the treatment on positive reviews resulting from review moderation transparency, this would further support the identified *reduction in contribution investment* as the main mechanism.

We first examined the impact of providing review moderation transparency on the volume of positive and negative ratings. We found that the treatment resulted in a decrease in the volume of positive ratings (Column 2 of Table 4) by 1.615 ($p < 0.001$) but only a marginal increase in the volume of negative ratings (Column 1 of Table 4) by 0.048 ($p < 0.1$). These results, shown in Table 4, suggest that providing review moderation transparency impacts review volume by minimizing the effort users invest in contributing positive reviews, which lends support to the identified mechanism of a reduction in investment contribution.

Second, we examined the differential impact of review moderation on the length of positive and negative reviews. As argued earlier, when users craft positive reviews, they may be concerned that their reviews might appear similar to fraudulent reviews when viewed from the perspective of the review filtering system. To avoid the wasted effort of writing elaborate reviews that could be filtered out, users may write less or more in their reviews, depending on the mechanism at work. Such concerns are likely minimal when users write negative reviews because such reviews are less likely to be filtered out. We expected a more significant decrease in the length of positive reviews than that of negative reviews in the presence of review moderation transparency if the decreased investment in review contribution is the primary underlying mechanism shaping user behaviors. Alternatively, if the increased contribution diligence mechanism is dominant, we would expect the review length of positive and negative reviews to increase rather than decrease, as users would want to provide more details to prove that their reviews are not fake. Specifically, we measured the average length of reviews with only negative sentiments and those with only positive sentiments for each month and then estimated the DID models reflected in Equation (1). Reviews with negative sentiments are reviews whose sentiment scores are lower than 0.5, and reviews with positive sentiments are reviews whose sentiment scores are greater than or equal to 0.5. The results in Columns 3 and 4 of Table 4 show that review moderation transparency reduced the length of both types of reviews, but the effect was greater for reviews with positive sentiment (Column 4), suggesting reduction in contribution investment as the most likely mechanism.

$$y_{itp} = \beta_2 High\_overall_i \times Post_t \times Treat_p + \beta_1 Post_t \times Treat_p + High\_overall_i \times Post_t + X_{itp} + \alpha_{ip} + \gamma_t + \varepsilon_{itp} \qquad (2)$$

To further understand the dominant mechanism, we performed an additional analysis using Equation (2) above to examine the impact of the treatment on online ratings for businesses for which the prior overall rating was high. A prior high rating could signal a greater potential that reviews would be filtered out, as most reviews were positive for those businesses. Thus, it is more likely that these businesses would have fake 5-star reviews, compared to businesses with lower prior overall ratings. Therefore, users might suspect that review filters would pay more attention to these businesses. If users believe that positive reviews of such businesses are more likely to be filtered out, they may be less motivated to write positive reviews, which could result in online ratings becoming more negative for businesses with higher prior ratings. Prior overall ratings were measured by accumulative average prior ratings at the system change; *high_overall* is a dummy indicating if a restaurant's prior overall rating was above average in our sample. The results in Column 5 of Table 4 show that online ratings were more significantly reduced (β = -0.202, $p < 0.001$) for businesses with high prior overall ratings compared to businesses with low prior overall ratings, supporting the mechanism *reduction in contribution investment*.

The Parallel Trends Assumption Test Results

Following Angrist and Pischke (2008) and many IS scholars (Chen et al., 2018; Greenwood & Wattal, 2017; Huang et al., 2017), we empirically assessed the parallel trend assumption. Specifically, we assessed the differences between the treatment and control groups in our outcome variables around the time of treatment by interacting a platform dummy with monthly time dummies. In order to establish the appropriateness of our control group, the difference between treated and untreated restaurants before treatment should be insignificant. The dynamic DID specification we used to model y_{it} is illustrated in Equation (3).

$$y_{itp} = \mu' \varphi_t \times Treat_p + \tau_t + X_{itp} + \alpha_{ip} + \varepsilon_{\ itp}, \qquad (3)$$

where most of the variables' definitions are the same as in Equation (1). γ_t represents time fixed effects for each quarter before and after Yelp's system change, μ' is a vector of coefficients to be estimated for the interaction terms between time dummies φ_t (i.e., the relative chronological distance between time *t* and the time of Yelp's system change) and the treatment dummy variable $Treat_p$, and α_{ip} denotes restaurant fixed effects.

The results from the relative-time model in Table 5 show that, with the exception of review length, the majority of the pretreatment time dummies are not significant. The results allowed us to validate that the pre-treatment trends are generally parallel. In terms of review volume, there are three significant coefficients in the pre-treatment period. However, there are two significant coefficients at the beginning of the pretreatment period, and then coefficients become insignificant, which does not support a preexisting trend. Second, in terms of review length, some of the pre-treatment time dummies are significant. However, the pre-treatment dummies closer to the treatment are insignificant. The post-treatment trend is consistently lower than the pre-treatment trend in the difference-in-differences, suggesting we have identified the true treatment effect.

Table 5. Relative Time Model of Review Moderation Transparency on Review Volume and Review Characteristics

Variables	(1) Review volume	(2) Ln (review length)	(3) Negative sentiment	(4) Review valence
Rel time(t-8)	-1.530***(0.393)	1.011***(0.072)	0.034(0.024)	0.135(0.086)
Rel time(t-7)	-1.368***(0.335)	0.955***(0.069)	0.037(0.023)	0.019(0.084)
Rel time(t-6)	-0.508(0.299)	0.738***(0.081)	0.028(0.025)	0.004(0.091)
Rel time(t-5)	0.413(0.279)	0.863***(0.079)	0.025(0.024)	0.065(0.087)
Rel time(t-4)	-0.561(0.306)	0.223**(0.070)	0.038(0.023)	-0.178*(0.086)
Rel time(t-3)	-0.027(0.301)	0.190**(0.063)	0.005(0.025)	-0.146(0.086)
Rel time(t-2)	-0.490(0.255)	0.079(0.061)	0.028(0.024)	-0.102(0.085)
Rel time(t-1)	0.606*(0.250)	0.055(0.058)	0.005(0.023)	-0.092(0.085)
Rel time(t0)	Omitted base case			
Rel time(t+1)	0.355(0.247)	0.081(0.055)	-0.003(0.024)	-0.058(0.082)
Rel time(t+2)	0.367(0.298)	0.073(0.055)	-0.009(0.023)	-0.022(0.080)
Rel time(t+3)	1.559***(0.294)	0.026(0.057)	-0.019(0.024)	0.008(0.085)
Rel time(t+4)	1.239***(0.301)	-0.006(0.056)	-0.025(0.022)	0.041(0.085)
Rel time(t+5)	-1.099*(0.439)	-0.064(0.052)	0.039+(0.022)	-0.180*(0.078)
Rel time(t+6)	-5.846***(0.672)	-0.026(0.051)	0.062**(0.021)	-0.316***(0.074)
Rel time(t+7)	-9.145***(0.798)	0.202***(0.051)	0.077***(0.020)	-0.390***(0.074)
Rel time(t+8)	-11.352***(0.847)	0.192***(0.050)	0.076***(0.020)	-0.379***(0.072)
Observations	22293	22293	22293	22293
R-squared	0.797	0.43	0.198	0.316
Number of restaurants	1016	1016	1016	1016
Restaurant fixed effect	Yes	Yes	Yes	Yes
Time fixed effect	Yes	Yes	Yes	Yes
Restaurant controls in Table 1	Yes	Yes	Yes	Yes

Robustness Checks

To ensure empirical rigor, we conducted a number of analyses to establish the robustness of our estimates and to eliminate confounding factors that could have affected our results. First, review volume and review length in our study are count variables. In such cases, a negative binomial model is often superior in analyzing response variables because it uses the correct probability distributions. Therefore, we estimated our results using negative binomial regression and reported the results in Column 1 and Column 2 of Table 6; the estimates are consistent with our main findings.

Second, if Yelp intensified its efforts to screen reviews after the system change or became more tolerant or strict with suspicious reviews, this practice would contribute to the volume change during the post-treatment period. However, we can discount the explanation that the decreased review volume on Yelp was due to more filtering after the treatment for several reasons. First, Luca and Zervas (2016) found that Yelp consistently filtered approximately 16% of reviews each year from 2008-2012. Second, one of the system changes addressed by our study is the disclosure of reviews that have been filtered out to users in order to gain trust. If Yelp filtered out more reviews after the system change, this would be observable to users and potentially erode user trust. Because of this, it is unlikely that Yelp would choose to combine transparency with increased scrutiny in content filtering. Third, by checking Yelp's news page, we explored whether Yelp fundamentally changed its filtering system during our observational period but found no evidence of this.

Third, as our sample is composed of the reviews that passed the filtering system and we assume that Yelp did an adequate job of flagging problematic reviews, the composition of our sample should not be affected by Yelp's improvement in flagging fraudulent reviews or changes in the behavior of fake reviewers. In support of this position, Luca and Zervas (2016) found that Yelp consistently filtered approximately 16% of reviews each year from 2008-2012 and that Yelp filters reviews of different ratings rather than just negative or positive reviews.

Table 6. Estimation Using Negative Binomial Regression

Variables	(1) Review volume	(2) Review length
Post × Treatment	-0.235***(0.033)	-0.217***(0.023)
Observations	41176	41176
Pseudo *R*-squared	0.21	0.015
Number of restaurants	1016	1016
Restaurant fixed effect	Yes	Yes
Time fixed effect	Yes	Yes
Restaurant controls in Table 1	Yes	Yes

Table 7. Effect of Review Moderation Transparency on Review Volume and Review Characteristics (Observation Period=36 Months)

Variables	(1) Review volume	(2) Negative sentiment	(3) Review valence	(4) Ln (review length)
Post × Treatment	-2.651***(0.248)	-0.485***(0.025)	0.024***(0.006)	-0.261***(0.022)
Observations	64752	64752	64752	64752
R-squared	0.716	0.296	0.117	0.223
Number of restaurants	1136	1136	1136	1136
Restaurant fixed effect	Yes	Yes	Yes	Yes
Time fixed effect	Yes	Yes	Yes	Yes
Controls in Table 1	Yes	Yes	Yes	Yes

Fourth, if other system changes unrelated to content filtering were implemented by either TripAdvisor or Yelp during our observational period, it may have also accounted for some of the observed differences in the outcome variables. We found that the only major changes that may have affected the validity of our results are Yelp's integration with social media accounts on July 2, 2009 (Huang et al., 2017) and TripAdvisor's adoption of a multidimensional rating system in January 2009 (Chen et al., 2018). Such integration has been found to increase review volume and reduce review negativity. Adopting a multidimensional rating system can reduce review negativity by increasing product ratings, but our finding is that providing review moderation transparency decreases review volume and increases review negativity. If social integration and the adoption of a multidimensional rating system were the dominant factors influencing review volume and negativity in our setting, we would expect to find reduced review negativity and increased review volume. In addition, we extended our observation window to 36 months to further establish the robustness of our results. The results summarized in Table 7 are generally consistent with our main findings and are inconsistent with those of Huang et al. (2017) and Chen et al. (2018). As a result, concerns that the difference in our outcome variables is partially or totally driven by the social integration and multidimensional rating features or other system changes that occurred near our event time should be greatly alleviated.

Finally, following Bertrand et al. (2004), we performed an additional robustness test to address concerns relating to standard errors caused by serial correlation within the residuals of difference-in-difference estimations. We ran a random implementation model to remove the concern that observed differences in our response variables are entirely by chance. First, we randomly applied the treatment to half of our sample of 1,016 restaurants, regressed the dependent variables on this pseudo treatment, and stored the coefficient. We replicated the regression 500 times for each dependent variable and compared the actual treatment's coefficient with the mean and standard deviation of the pseudo-treatment coefficient. The results in Table 8 show that all the coefficients obtained by randomizing treatments significantly differ from those in the main analysis. Therefore, it is highly unlikely that the observed coefficients appeared purely by chance.

Table 8. Output of Random Implementation Model of Treatment

Variables	(1) Review volume	(2) Ln (review length)	(3) Negative sentiment	(4) Review valence
μ of Random β	0.057	-0.120	-0.004	-0.006
σ Random β	0.213	0.018	0.005	0.022
Estimated β	-1.786	-0.416	0.012	-0.181
Replications	500	500	500	500
Z-score	-8.650	-16.674	3.074	-7.854
P-value	$p < 0.001$	$p < 0.001$	$p < 0.001$	$p < 0.001$

Discussion

Key Findings

Though the importance of information transparency in B2C markets has been widely acknowledged, significant gaps exist in examining information transparency in online review platforms, which have become an integral part of the shopping experience for an increasingly large segment of the economy. Our study investigates the effects of providing review moderation transparency on review volume, length, and negativity. Specifically, we found that providing review moderation transparency reduced the volume and length of reviews, two critical dimensions of reviews that capture users' investments in review contribution.

In addition, we find increased review moderation transparency resulted in increased negativity—i.e., an increase in the negative sentiment of reviews and decreased online ratings. Considering the findings in the literature that users perceive negative and longer reviews to be more helpful (Cao et al., 2011; Chen & Lurie, 2013; Kuan et al., 2015; Mudambi & Schuff, 2010), our results imply that providing review moderation transparency decreases user investment in review contribution but has a complex effect on the helpfulness of reviews contributed to online platforms. On the one hand, helpfulness is enhanced because online ratings and average review sentiment are more negative. On the other hand, helpfulness is reduced due to shorter reviews. In addition, we also found that the presence of review moderation transparency resulted in (1) a decline in the volume of positive ratings, (2) a greater decline in the length of reviews with positive sentiment than reviews with negative sentiment, and (3) a greater reduction in online ratings for businesses that had higher prior overall ratings. These findings suggest that the mechanism associated with the impact of review moderation transparency on the outcomes is likely driven by a reduction in contribution investment, which has implications for the characteristics of reviews contributed. These nuanced effects mean that platforms pay a price for increased transparency in review moderation. While the increased transparency may help to address user concerns regarding advertiser preference in review filtering, a healthy review platform requires substantial user content submissions to ensure that the platform remains an essential tool for decision makers.

Limitations and Future Work

As with any research, our study has some limitations. First, we tried to rule out the confounding impact of the extent of review moderation implemented by the platform through various robustness checks. Nonetheless, since we cannot examine each online platform's review moderation mechanisms directly, shifts in the extent of review moderation are confounding factors that cannot be ruled out. This could be explored in future work involving direct collaboration with review platforms.

Second, we treated the two Yelp platform changes as one system change, as the dates of providing video explanations and showing filtered reviews are very close to each other. Future studies could explore the design changes of different platforms or use experiments to investigate the relative role of video explanations vs. showing filtered reviews on platform outcomes.

Third, we tried to tease out the primary underlying mechanism between providing review moderation transparency and platform outcomes. However, there are also shifts in behavior that we could not observe, such as the behaviors of individuals submitting fake reviews. Future work should consider the impact of transparency on the behavior of individuals submitting fake reviews and review filter quality via either controlled lab experiments or collaborations with review platforms. Finally, we assume that consumers writing reviews on Yelp were also paying attention to the system change providing review moderation transparency. Future research could examine whether users actually observed this review moderation transparency policy. It is also possible that some consumers were indifferent to the system change, reducing the likelihood of us finding significant effects in our outcome variable.

Contributions to Research

This research makes several contributions to the literature. First, the present study represents one of the first efforts at quantifying the impact of platform-level content moderation transparency on the contribution of content in the context of online reviews. Although prior studies have called for papers on how firms' design policies for selective information disclosure affect users (Granados & Gupta, 2013), few empirical studies have examined the impact of providing content moderation transparency. The novel findings reported here show how providing content moderation transparency can more broadly influence contributions to online platforms, even if users do not directly experience their content being moderated using online review data.

Second, our study extends the emerging transparency stream in IS research by examining the nuanced impacts of providing information transparency. Much of the current research in this area has examined the positive consequences of information transparency and argued for the need to provide more transparency (e.g., Granados et al., 2006; Schnackenberg & Tomlinson, 2016; Tapscott & Ticoll, 2003). In addition, providing content moderation transparency has been promoted in the "Santa Clara Principles on Transparency and Accountability Around Content Moderation" (York, 2020), but the impacts of such transparency on user behavior on platforms are unclear. We contribute to this research area and showcase the importance of investigating content moderation transparency in the context of online reviews because of the nuanced impacts of providing transparency. Overall, we found evidence that an increase in transparency results in an associated *reduction in contribution investment* among users, resulting in shorter, fewer, and more negative reviews. Since shorter reviews are perceived to be less helpful by prior studies (Mudambi & Schuff, 2010), providing review moderation transparency may negatively affect information quality on online platforms. However, induced higher review negativity suggests that review moderation transparency may also boost review helpfulness on online platforms, as review negativity is positively related to review helpfulness (Cao et al., 2011; Chen & Lurie, 2013). Our paper thus sheds light on the nuanced effects of review moderation transparency as a design choice for online platforms.

Third, in addition to expanding our understanding of information transparency, our study contributes to the WOM literature. We extend this work by providing evidence of how review moderation—which reminds users about the possibility of platform-level censorship and sanctions—influences user behaviors on online platforms. This mechanism is quite different from those examined in prior studies that have focused on social evaluation from a socially proximal audience such as friends or family members (e.g., Burtch et al., 2015; Huang et al., 2017). Our findings suggest that evaluation by socially distant authority entities might shape user behaviors very differently. This has several implications that require careful examination in future studies as platforms and government agencies contemplate content moderation policies.

Practical Implications

This study provides critical practical guidelines for firms developing, hosting, and marketing online platforms. While online review platforms may disclose information about the review moderation process to gain user trust and address concerns of preferential filtering for advertisers, they should also consider the negative impact of implementing this policy. Users might be told that their reviews will be evaluated, and an associated change in contribution may follow, reducing the likelihood of users contributing reviews and inhibiting longer reviews. As a result, firms should consider adopting proactive measures to mitigate such negative impacts. For example, platforms could reinforce the value of screening out inappropriate content and reassure users that sharing authentic feelings about their experience with a service will not be censored. In addition, platforms could provide reassuring metrics emphasizing how infrequently content is removed from the platform. However, we also found that transparency facilitates using negative words in reviews and producing negative ratings. Due to negativity bias (Rozin & Royzman, 2001), users might find platforms with more negative content in reviews to be more diagnostic when attempting to perform pre-purchase analysis and thus may engage more with the platform and become more loyal to it. Platforms are well advised to carefully consider such consequences when implementing review moderation and making it transparent.

Prior studies have suggested that increases in review volume and ratings on an online platform boost sales (Liu, 2006; Luca & Zervas, 2016). Thus, firms listed on a transparent platform might experience an unfavorable shift in their ratings and user sentiment, not because of any changes in their products or services but because of the platform's design choice. Platforms striving to be transparent must employ means of countering such unintended consequences. One possible way to deal with this could be that transparent platforms could cue users to the average characteristics of the reviews on their platform for comparable products or services and allow an outside firm to benchmark its ratings and reviews.

Concluding Remarks

Given the sparsity of theoretically anchored frameworks to understand platform design choices, such as moderation transparency on user behaviors, this paper fills a critical gap. It provides a rich understanding of the mechanism through which transparent content moderation shapes the content-contributing behaviors of users and the potential consequences of such behavior shifts. In addition, this paper guides platforms on how to implement content moderation policies and deal with the multifaceted effects that such platform design choices engender.

Acknowledgments

The authors express their gratitude to the senior editor, H. Raghav Rao, and the associate editor, for their enduring patience, constructive feedback, thoughtful guidance, and tireless support. The authors are also grateful for the time and effort of three anonymous reviewers, whose critiques have improved this paper considerably.

References

Angrist, J. D., & Pischke, J.-S. (2008). *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press.

Bertrand, M., Duflo, E., & Mullainathan, S. (2004). How much should we trust differences-in-differences estimates? *The Quarterly Journal of Economics*, *119*(1), 249-275. https://doi.org/10.1162/003355304772839588

Boberg, S., Schatto-Eckrodt, T., Frischlich, L., & Quandt, T. (2018). The moral gatekeeper? Moderation and deletion of user-generated content in a leading news forum. *Media Communication*, *6*(4), 58-69. https://doi.org/10.17645/mac.v6i4.1493

Burtch, G., Ghose, A., & Wattal, S. (2015). The hidden cost of accommodating crowdfunder privacy preferences: A randomized field experiment. *Management Science*, *61*(5), 949-962. https://doi.org/10.1287/mnsc.2014.2069

Cao, Q., Duan, W., & 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. https://doi.org/10.1016/j.dss.2010.11.009

Chen, P.-Y., Hong, Y., & Liu, Y. (2018). The value of multidimensional rating systems: Evidence from a natural experiment and randomized experiments. *Management Science*, *64*(10), 4629-4647. https://doi.org/10.1287/mnsc.2017.2852

Chen, Z., & Lurie, N. H. (2013). Temporal contiguity and negativity bias in the impact of online word of mouth. *Journal of Marketing Research*, *50*(4), 463-476. https://doi.org/10.1509/jmr.12.006

Chevalier, J. A., & Mayzlin, D. (2006). The effect of word of mouth on sales: Online book reviews. *Journal of Marketing Research*, *43*(3), 345-354. https://doi.org/10.1509/jmkr.43.3.34

Cook, C. L., Patel, A., & Wohn, D. Y. (2021). Commercial versus volunteer: Comparing user perceptions of toxicity and transparency in content moderation across social media platforms. *Frontiers in Human Dynamics*, *3*, Article 3. https://doi.org/10.3389/fhumd.2021.626409

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

Erskine, R. (2017). Yelp's "don't ask" policy is bad for everyone ... including Yelp. *Forbes.* https://www.forbes.com/sites/ryanerskine/2017/11/16/yelps-dont-ask-policy-is-bad-for-everyone-including-yelp/#353bdc93103d

European Commission. (2018). *Behavioural study on transparency in online platforms*. https://ec.europa.eu/info/publications/behavioural-study-transparency-online-platforms-2018_en

Ghose, A., & 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. https://doi.org/10.1109/TKDE.2010.188

Goes, P. B., Lin, M., & Au Yeung, C.-m. (2014). "Popularity effect" in user-generated content: evidence from online product reviews. *Information Systems Research*, *25*(2), 222-238. https://doi.org/10.1287/isre.2013.0512

Gosling, S. D., Augustine, A. A., Vazire, S., Holtzman, N., & Gaddis, S. (2011). Manifestations of personality in online social networks: Self-reported Facebook-related behaviors and observable profile information. *Cyberpsychology, Behavior, and Social Networking*, *14*(9), 483-488. https://doi.org/10.1089/cyber.2010.0087

Granados, N., & Gupta, A. (2013). Transparency strategy: Competing with information in a digital world. *MIS Quarterly*, *37*(2), 637-641.

Granados, N. F., Gupta, A., & Kauffman, R. J. (2006). the impact of it on market information and transparency: A unified theoretical framework. *Journal of the Association for Information Systems*, *7*(3), 148-178. https://doi.org/10.17705/1jais.00083

Greenwood, B. N., & Wattal, S. (2017). Show me the way to go home: An empirical investigation of ride-sharing and alcohol related motor vehicle fatalities. *MIS Quarterly*, *41*(1), 163-187.

Hu, N., Pavlou, P. A., & Zhang, J. (2006). Can online reviews reveal a product's true quality? Empirical findings and analytical modeling of online word-of-mouth communication. In *Proceedings of the 7th ACM Conference on Electronic Commerce*.

Hu, N., Zhang, J., & Pavlou, P. A. (2009). Overcoming the J-shaped distribution of product reviews. *Communications of the ACM*, *52*(10), 144-147. https://doi.org/10.1145/1562764.1562800

Huang, N., Hong, Y., & Burtch, G. (2017). Social network integration and user content generation: Evidence from natural experiments. *MIS Quarterly*, *41*(4), 1035-1058. https://doi.org/10.25300/MISQ/2017/41.4.02

Jhaver, S., Bruckman, A., & Gilbert, E. (2019). Does transparency in moderation really matter? User behavior after content removal explanations on Reddit. In *Proceedings of the ACM on Human-Computer Interaction*. https://doi.org/10.1145/3359252

Juneja, P., Rama Subramanian, D., & Mitra, T. (2020). Through the looking glass: Study of transparency in Reddit's moderation practices. In *Proceedings of the ACM on Human-Computer Interaction*. https://doi.org/10.1145/3375197

Klepper, D., & O'Brien, M. (2021). *Social platforms flex their power, lock down Trump accounts*. AP. https://apnews.com/article/facebook-ban-trump-3e9a00e791f9806a4d925ec9a2fbe9f3

Kuan, K. K., Hui, K.-L., Prasarnphanich, P., & Lai, H.-Y. (2015). What makes a review voted? An empirical investigation of review voting in online review systems. *Journal of the Association for Information Systems*, *16*(1), 48-71. https://doi.org/10.17705/1jais.00386

Liu, Y. (2006). Word of Mouth for Movies: Its dynamics and impact on box office revenue. *Journal of Marketing*, *70*(3), 74-89. https://doi.org/10.1509/jmkg.70.3.0

Luca, M., & Zervas, G. (2016). Fake it till you make it: Reputation, competition, and Yelp review fraud. *Management Science*, *62*(12), 3412-3427. https://doi.org/10.1287/mnsc.2015.2304

Moe, W. W., & Schweidel, D. A. (2012). Online product opinions: incidence, evaluation, and evolution. *Marketing Science*, *31*(3), 372-386. https://doi.org/10.1287/mksc.1110.0662

Mudambi, S. M., & Schuff, D. (2010). What makes a helpful online review? A study of customer reviews on Amazon.com. *MIS Quarterly*, *34*(1), 185-200.

Oh, W., & Lucas Jr, H. C. (2006). Information technology and pricing decisions: Price adjustments in online computer markets. *MIS Quarterly*, *30*(3), 755-775.

Patterson, B. (2016). *5 Yelp facts business owners should know (but most don't)*. MarTech. https://marketingland.com/5-yelp-facts-business-owners-should-know-163054

Prentice-Dunn, S., & Rogers, R. W. (1982). Effects of public and private self-awareness on deindividuation and aggression. *Journal of Personality and Social Psychology*, *43*(3), 503-513. https://doi.org/10.1037/0022-3514.43.3.503

Rabin, M. (2013). Risk aversion and expected-utility theory: a calibration theorem. In L. C. MacLean & W. T. Ziemba (Eds.), *Handbook of the fundamentals of financial decision making: Part I* (pp. 241-252). World Scientific. https://doi.org/10.1142/9789814417358_0013

Rozin, P., & Royzman, E. B. (2001). Negativity bias, negativity dominance, and contagion. *Personality and Social Psychology Review*, *5*(4), 296-320. https://doi.org/10.1207/S15327957PSPR050

Schnackenberg, A. K., & Tomlinson, E. C. (2016). Organizational transparency: A new perspective on managing trust in organization-stakeholder relationships. *Journal of Management*, *42*(7), 1784-1810. https://doi.org/10.1177/0149206314525202

Shane, S. (2018). Five takeaways from new reports on Russia's social media operations. *The New York Times*. https://www.nytimes.com/2018/12/17/us/politics/takeaways-russia-social-media-operations.html

Shulman, J. D., Cunha Jr., M., & Saint Clair, J. K. (2015). Consumer uncertainty and purchase decision reversals: Theory and evidence. *Marketing Science*, *34*(4), 590-605. https://doi.org/10.1287/mksc.2015.0906

Snyder, B. (2015). Yelp says FTC has dropped inquiry into its reviews. *Fortune*. http://fortune.com/2015/01/06/yelp-ftc-inquiry/

Solis, N. (2018). *Why does Yelp filter reviews & what you can do about it*. Broadly. https://broadly.com/blog/why-does-yelp-filter-reviews/

Sparks, B. A., & Browning, V. (2011). The impact of online reviews on hotel booking intentions and perception of trust. *Tourism Management*, *32*(6), 1310-1323. https://doi.org/10.1016/j.tourman.2010.12.011

Suzor, N. P., West, S. M., Quodling, A., & York, J. (2019). What do we mean when we talk about transparency? Toward meaningful transparency in commercial content moderation. *International Journal of Communication*, *13*, 1526-1543.

Tapscott, D., & Ticoll, D. (2003). *The naked corporation: How the age of transparency will revolutionize business*. Simon & Schuster.

Trifts, V., & Häubl, G. (2003). Information availability and consumer preference: Can online retailers benefit from providing access to competitor price information? *Journal of Consumer Psychology*, *13*(1-2), 149-159. https://doi.org/10.1207/S15327663JCP13-1&2_13

Vaughan, B. (2020). Don't just boycott Facebook: Create something better! *Forbes*. https://www.forbes.com/sites/benjaminvaughan/2020/07/07/I-just-boycott-face—ok--create-something-better/#7df486a84d3d

Wang, W., & Benbasat, I. (2007). Recommendation agents for electronic commerce: Effects of explanation facilities on trusting beliefs. *Journal of Management Information Systems*, *23*(4), 217-246. https://doi.org/10.2753/MIS0742-1222230410

Wang, W., & Benbasat, I. (2016). Empirical assessment of alternative designs for enhancing different types of trusting beliefs in online recommendation agents. *Journal of Management Information Systems*, *33*(3), 744-775. https://doi.org/10.1080/07421222.2016.1243949

Yelp. (2010). *Yelp's recommendation software explained*. https://www.yelpblog.com/2010/03/yelp-review-filter-explained

York, J. (2020). *What comes next for the Santa Clara Principles: 2020 in Review*. EFF. https://www.eff.org/deeplinks/2020/12/2020-year-review-what-comes-next-santa-clara-principles

About the Authors

Lianlian (Dorothy) Jiang is an assistant professor in the Department of Decision & Information Sciences at the Bauer College of Business, University of Houston. She has published work in *MIS Quarterly*, *Communications of the Association for Information Systems* and *3D Research*, as well as in proceedings from the Americas Conference on Information Systems, the *International Conference on Information Systems*, and others. Her research interests are in the areas of digital platform design & strategy, healthcare IT, human-AI interaction, and business analytics.

T. Ravichandran is the Associate Dean for Research and the Irene and Robert Bozzone'55 Distinguished Professor of Management & Technology in the Lally School of Management, Rensselaer Polytechnic Institute. His research has been published or forthcoming in journals such as the *Communications of the ACM, Decision Sciences, European Journal of Information Systems, IEEE Transactions on Engineering Management, Information Systems Research, Information Technology and Management, Journal of Management Information Systems, Journal of Strategic Information Systems, Journal of Service Research, MIS Quarterly* and *Organization Science*. His research is widely cited and has won several awards, including the Association of Information Systems Best Information Systems Publication Award in 2010 and the *Information Systems Research* Best Published Paper Award in 2010.

His research has been supported by grants from the National Science Foundation and from the Ministry of Education, Singapore. He has served as a Senior Editor of *MIS Quarterly*, as a Department Editor of IEEE Transactions in Engineering Management, and as an Associate Editor of both *MIS Quarterly* and *Information Systems Research*. He works closely with large companies and startups on digital strategy, innovation. and supply chain management and is a frequent speaker in many industry and academic forums around the world.

Jason Kuruzovich is an Associate Professor of Business Analytics at Rensselaer Polytechnic Institute (RPI). Jason is an impact-driven researcher who studies online platforms and the implications resulting from their role in society. Current streams of research investigate the role of AI in hiring, gaming, and electronic commerce. His work has been published in top journals in information systems, marketing, and organizational behavior.