I Hear You:
Does Quality Improve with Customer Voice?*

Uttara Ananthakrishnan†
Foster School of Business, University of Washington
uttara@uw.edu

Davide Proserpio
Marshall School of Business, University of Southern California
proserpi@marshall.usc.edu

Siddhartha Sharma
Kelley School of Business, Indiana University
sishar@iu.edu

March 2, 2023

*The authors thank Brett Danaher, Anthony Dukes, Shantanu Dutta, Michael Luca, Leonardo Madio, Ananya Sen, Param Vir Singh, Michael Smith, Rahul Telang, and Gerard Tellis for valuable comments. This paper has also benefited from the comments of seminar participants at University of Texas at Austin, Indiana University, and Chinese University of Hong Kong, and attendees of 2018 Workshop on Information Systems and Economics, 2019 Conference on Information Systems and Technologies, 2019 Conference on Artificial Intelligence, Machine Learning and Business Analytics, 2020 Marketing Science Conference, 2020 ZEW Conference on the Economics of ICT, 2020 ACM Conference on Economics and Computation, 2021 NBER Summer Institute IT and Digitization Workshop, and 2022 Statistical Conference in E-Commerce Research. The authors also thank Google Cloud Platform for their generous support and for providing us with research credits to conduct some of the analyses reported in this paper, and STR for providing hotel performance data. All remaining errors are our own.

†Authors' names are listed in alphabetical order.

Abstract

In a static quality context, online reviews and ratings help consumers separate high- and low-quality firms. In a dynamic quality context, however, reviews can inform and incentivize low-rated firms to improve their quality and lower the quality gap with high-rated firms. In this paper, we empirically test this hypothesis by analyzing the U.S. hotel industry using data from two major online consumer review platforms: Tripadvisor and Expedia. Using a combination of econometric and natural language processing tools, we present the following findings. First, hotels that are more likely to pay attention to reviews increase their ratings more than hotels that are less likely to pay attention to reviews. Second, these hotels increase their ratings by improving on issues frequently mentioned in their reviews. Third, we find that low-rated hotels experience larger gains in ratings as they have more margin for improvement than high-rated hotels. Overall, our results suggest that online reviews are a valuable source of information for firms and may improve consumer welfare.

1 Introduction

Understanding consumer needs is at the core of marketing practice. Historically, marketing managers across industries relied on focus groups, experiential interviews, or ethnography as inputs to improve their product or service operations. Longstanding marketing literature discusses methods to identify consumer needs from feedback (Griffin and Hauser 1993, Alam and Perry 2002). However, with the advent of the Internet and the proliferation of online platforms, obtaining consumer feedback suddenly became much easier and democratized. Review platforms like Tripadvisor and Yelp have emerged as popular channels through which consumers express their opinions and provide feedback to firms. While online customer reviews have been shown to significantly impact purchase decisions (Luca 2016), it is unclear if firms use online customer reviews to improve their quality. This is the focus of our study.

Compared to privately collected offline feedback, online customer reviews provide a fundamentally different form of consumer opinion and incentives for firms to act on. First, online review platforms are public forums where consumers can voice their concerns. The public nature makes the feedback visible to future customers, unlike the traditional mode of obtaining privately held feedback. This increased transparency not only influences other customers and the feedback they may leave (Godes and Silva 2012) but could provide a stronger incentive for firms to improve their quality, as repeatedly ignoring reviews could disproportionately magnify the scale and severity of the issues raised by the customers, reducing the trust in the firm. Second, unlike surveys and focus groups where firms actively seek this feedback to learn more about their consumers, firms receive feedback from their customers through the review platforms regardless of whether they choose to obtain it. A firm facing this unsolicited feedback can decide whether to listen and act on the feedback or simply ignore it deeming it irrelevant or not a priority. Given these differences, it is critical to understand the role of online reviews in improving firm quality. In this paper, we empirically investigate whether and which firms improve the quality of their service based on the feedback provided by customers via online reviews. We study this question in the context of the U.S. hotel industry by combining econometric methods with state-of-the-art machine learning (ML) tools for natural language processing (NLP).

While listening to online consumer feedback could lead to quality improvements for all firms, substantial heterogeneity across firms can either facilitate or hinder the ability to improve the quality of their offering. For example, firms with many issues (and therefore with lower average ratings) might improve more than firms with fewer or no problems (and therefore with higher average ratings) because the former set of firms has a bigger room for improvement. In addition, firms with better resources and capability, e.g., chain or upscale hotels, might be more likely to extract information from reviews and take action than less resourceful firms like independent or budget hotels. Similarly, firms whose consumer reviews are more homogeneous, i.e., concentrated on only a few quality dimensions, might be more likely to take action than firms with significant heterogeneity in what customers complain about in the reviews. In summary, there could be several reasons why quality improvements could significantly vary across different types of firms that pay attention to online customer reviews.

We study the role of online reviews in quality improvement using data on consumer reviews and management responses from two major online platforms: Tripadvisor and Expedia. We focus on the hotel industry because it is more likely to observe feedback-based quality improvements in service-based industries than in physical product-based industries. In other words, quality is more flexible in the service-based industries, making it easier for these firms to act on negative feedback. This suggests that in a static quality context, i.e., when the quality of an offering cannot change, online reviews help consumers separate between low- and high-quality sellers. However, in a dynamic quality context, i.e., when the quality of an offering can change over time, online reviews can also help reduce the quality gap between low- and high-quality firms, as low-quality firms can benefit more from online customer feedback than high-quality firms.

There are multiple challenges to answering this question using observational data. First, obtaining a good, unbiased measure of hotel quality is hard. Second, whether a hotel pays attention to reviews is unobserved to us as researchers. Third, any improvement in quality that is observed could be associated with factors other than reviews. We overcome these challenges in the following way. First, we measure changes in quality using not only the star rating of the reviews but also the sentiment of the problematic entities mentioned in the reviews of a hotel. Second, we exploit variations in how hotels respond to reviews as a proxy for whether hotels are more likely to pay attention to reviews. More specifically, realizing online reviews' critical role in consumer decision-making, firms have started directly addressing the reviews with individual responses (Gu and Ye 2014, Proserpio and Zervas 2017). In these responses—particularly in those to negative reviews—firms often claim that they are (or will be) making improvements based on the issues highlighted in those reviews (Zhang and Vásquez 2014). We use management responses because past literature suggests that hotels use them to signal that they are listening to what consumers say.[1] Third, to isolate the impact of paying attention to reviews on hotels' quality, we employ a difference-in-differences (DD) identification strategy that exploits the fact that some hotels do not respond to reviews, and those that do respond started responding at different times. This type of generalized DD approach with variation in treatment timing has been previously used in several studies. It is an appealing technique to identify a control group when one is not readily available (Bertrand et al. 2004, Gentzkow et al. 2011, Burtch et al. 2018, Prager and Schmitt 2021).

Despite using the above identification strategy, we still face a significant challenge. Prior literature has shown that the presence of management responses can have important effects on consumers' reviewing behavior which, in turn, can affect firms' average star-ratings (Proserpio and Zervas 2017, Chevalier et al. 2018). In practice, this means that if we were to

[1] Chevalier et al. (2018) often state that because of management responses "reviewers receive a credible signal that the service provider is listening", while Proserpio and Zervas (2017) state that positive reviews increase because through management responses "the hotel has signaled that it is listening".

measure changes in ratings on a platform where responses to consumer reviews are available, our results would likely be biased by the mere presence of these responses. To isolate the impact of quality improvements on ratings, we exploit the fact that while both Tripadvisor and Expedia allow firms to respond to reviews, firms usually do so on Tripadvisor but rarely on Expedia. Our DD strategy uses a hotel's timing of their first response on Tripadvisor as the treatment variable, but its Expedia star ratings as the outcome variables.[2]

Using the above identification strategy, we present two findings suggesting that hotels do use reviews to improve their quality. First, we show that hotels that are more likely to listen to reviews improve their rating by 0.04 stars on average after they begin to respond. In addition, we find that this average increase in ratings is mainly driven by hotels that had poor ratings before responding to reviews, and these hotels see their ratings increase by about 0.1 stars. Finally, we show that, among the low-rated hotels, chain and upscale hotels realize larger gains compared to independent and budget hotels, respectively. Second, we provide direct evidence that the improvement in ratings is due to quality improvements driven by the reviews. We use state-of-art NLP algorithms to (i) identify frequently mentioned entities in the reviews for each hotel (e.g., "breakfast", "staff") before the hotel's first response; and (ii) extract the sentiment for each one of these entities. We then focus on entities with negative average sentiment and analyze the evolution of the sentiment of these entities before and after a hotel begins responding to reviews. Following this approach, we show that, on average, the negative sentiment associated with frequently mentioned entities becomes neutral or positive after managers start responding to consumer reviews. Moreover, the sentiment improvement is higher for entities that are mentioned more often and reviews that receive a response (which are more likely to be read). These findings suggest that managers listen to reviews and make feedback-based quality improvements to fix issues frequently mentioned in the reviews. Third, we show that hotels that are likely to care less about what customers

[2] Analyzing Expedia reviews has another advantage. Unlike on Tripadvisor, where anyone can post a review, only those users that have booked the hotel through Expedia and also stayed at the hotel for at least one night can write a review on Expedia. This drastically reduces the presence of fake reviews on Expedia.

have to say, i.e., hotels that copy and paste the same response to different reviews, or that write canned responses, do not improve their ratings as much as hotels that carefully respond to reviews. These results withstand several identification and robustness checks, including accounting for city-specific and hotel-specific time-varying shocks, changes in management, renovations, competition, and hotel prices.

Our paper contributes to the growing literature on online reviews and reputation management by showing that online customer feedback can help firms make quality improvements. This suggests that hotels not only respond in words to online reviews but also with actions. We provide empirical evidence of this phenomenon by leveraging both quantitative (numerical ratings) and qualitative (sentiment of review attributes) measures of quality.

Our work has important implications for consumers, firms, and review platforms. First, we show how online user-generated content influences the quality of firms and leads to overall quality improvement in a service-based industry, potentially improving consumer welfare. Second, we show that online review platforms can create value for low-quality firms by informing them of their issues and allowing them to redeem themselves in the eyes of the customers, therefore increasing their ratings. Third, while past research has studied ways in which online reviews have benefited smaller independent firms (Hollenbeck 2018), our research shows that when it comes to quality improvement, larger firms like chain or upscale hotels gain the most from online customer reviews. Finally, when customers realize that firms listen to them and take action based on their feedback, they might be further incentivized to write reviews online, leading to a better representation of customer opinions on online platforms.

2 Related Literature

Learning from consumers. Mitsubishi's Kobe shipyard developed the concept of quality function deployment (QFD), which aids an inter-functional team in developing new products based on an assessment of customers' needs (Sullivan 1986). Since then, a long-standing body of literature has discussed various ways in which firms can learn from their consumers to improve product development (Krishnan and Ulrich 2001). Hauser et al. (1988) focus on the first stage of QFD, in which the voice of the customer is understood and then translated into the voice of engineers. Berry and Parasuraman (1997) introduce the service-quality information system, in which multiple methods are used to listen to different consumer groups. Matzler and Hinterhuber (1998) discuss how Kano's model of customer satisfaction can be integrated with QFD. Herrmann et al. (2000) talk about bridging the gap between identifying customer needs, managing quality, and measuring customer satisfaction.

Related to these papers, and beginning with Griffin and Hauser (1993), several papers have developed methods to identify customer needs. These methods rely on focus groups, experiential interviews, or ethnography as inputs. Professionals then analyze each input and manually identify and structure the customer needs (Kaulio 1998, Alam and Perry 2002). More recently, Schaffhausen and Kowalewski (2015) proposed using a web interface to collect customer needs and stories by asking customers directly, while Timoshenko and Hauser (2019) proposed a deep learning method to mine customer needs from user-generated content such as online consumer reviews. Finally, Bertschek and Kesler (2017) show that firms adoption of a Facebook page, as well as feedback from users, are positively and significantly related to product innovations. We extend this line of research to a dynamic quality environment by investigating whether service-based firms improve over time by resolving the issues raised by their customers on online review platforms.

Online reviews. While most of the papers in the online word-of-mouth literature have focused on the impact of reviews on the consumer side, minimal attention has been paid to the effect on firms' decision-making. Feng et al. (2019) show how online reviews can affect the pricing strategies of firms, whereas Wang et al. (2020) study how reduced information asymmetry between consumers and firms due to online reviews can incentivize independent hotels to improve their quality. Similar to this paper, we study whether and how customer reviews improve their quality.

The practice of responding to reviewers has been growing in popularity in the last few years, and many researchers started studying the effect of this practice on subsequent reviewing behavior of consumers (Proserpio and Zervas 2017, Chevalier et al. 2018, Wang and Chaudhry 2018, Proserpio et al. 2021). Importantly for our identification strategy, both Chevalier et al. (2018) and Proserpio and Zervas (2017) argue that the presence of management responses signals that the service provider is listening. Unlike these papers that analyze the mechanism by which management responses affect consumer ratings, we investigate whether responding to reviews suggests that hotels are improving their ratings by making actual quality improvements based on online feedback. A key difference between these papers and ours is that while they exploit cross-platform variation in hotel ratings, we exploit within-platform variation in hotel ratings. In doing so, we can completely eliminate the effect that the presence of management responses has on ratings (because we focus on a platform, Expedia, with no management responses during our observation period) and thus better isolate the association between paying attention to reviews and feedback-based quality improvements.

Text as data. With the rise of social media platforms over the past two decades, a massive amount of data is generated in the text format, leading many marketing researchers to analyze textual data (Berger et al. 2020). Several researchers (Lee and Bradlow 2011, Netzer et al. 2012, Büschken and Allenby 2016, Puranam et al. 2021) have developed methods to extract relevant product or brand attributes and topics from user-generated content. Many of these methods rely on dictionary-based techniques, word co-occurrences, or Latent Dirichlet Allocation. Following recent developments in computer science, more recent studies such as Adamopoulos et al. (2018), Liu et al. (2018), Chakraborty et al. (2021) and Hartmann et al. (2022) have used deep learning models that take into account the syntactic structure of sentences. These methods can extract text features that cannot be computed using traditional methods to analyze text data. Apart from mining topics from a text document, there has also been considerable work on studying the role of the sentiment of specific words or sentences (Tirunillai and Tellis 2012, Schweidel and Moe 2014, Mousavi and Gu 2019). Like these methods, we employ recently developed deep learning tools to analyze and extract features and sentiment from the text of the reviews and responses in our data.

3 Data

3.1 Tripadvisor and Expedia

Funded in 2000, Tripadvisor is one of the world's largest websites for travel-related reviews. As of 2019, Tripadvisor contains 760 million reviews of 8.3 million accommodations (hotels, B&Bs, inns), restaurants, experiences, airlines, and cruises. We collected reviews (and management responses) left on the website over 13 years (from 2001 to 2013) for all hotels available in the top 50 cities by population in the US. We also extract information on a hotel's operation (chain or independent) and class (low-, mid-, and high-tier).[3] The final Tripadvisor dataset contains 1,755,878 reviews for 7,318 hotels, and 510,706 responses from 4,815 hotels.

Expedia is one of the world's largest online travel agencies. Expedia allows users to plan and book their next vacation, from accommodations to flights and cars. Like Tripadvisor, Expedia also allows its users to review the service booked. An essential difference between the two platforms is that, while anyone can write a review on Tripadvisor, users can only write a review on Expedia if they bought (and consumed) the service through Expedia. In practice, this means that one is less likely to find fake reviews on Expedia than on Tripadvisor (Mayzlin et al. 2014). As we did for Tripadvisor, we collected the reviews (and their responses) of all hotels in the top 50 cities by population in the US for ten years, from September 2004 to

[3] The hotel class is defined using Tripadvisor's hotel class classification system.

December 2013. The final Expedia dataset contains 2,798,752 reviews for 5,570 hotels, out of which 1,192 hotels responded to 28,542 of their reviews.

3.2 Linking the Data and Sample Selection

We link each hotel on Expedia with its corresponding entry on Tripadvisor using the Tripadvisor link available on the hotel's entry on Expedia.[4] Using this procedure, we can link 5,354 hotels (96% of all Expedia hotels) between the two platforms. These 5,354 matched hotels have 2,749,049 reviews with a star rating on Expedia, out of which only 28,179 (or about 1%) have a response (these responses are written by 1,165 hotels and, on average, these hotels responded to 3% of their reviews). For the same set of hotels, there are 1,652,680 reviews from Tripadvisor, 500,811 (about 30%) of which received a response.

We exclude 1,165 of the matched 5,354 hotels from our analysis because their managers also responded to at least one review on Expedia before 2014. Excluding these hotels ensures that the Expedia ratings are not affected by changes in consumer review behavior as documented by Proserpio and Zervas (2017). The final Expedia dataset contains 4,189 hotels with 1,857,040 reviews, covering about ten years, from 2004 to 2013. The total number of reviews and responses for the same hotels on Tripadvisor amount to 1,145,155 and 298,617, respectively. We exclude 9 out of the 4,189 hotels that started responding before our sample period.

3.3 Descriptive Statistics

Of the $4,180$ hotels, $1,070$ hotels did not respond to any review on Tripadvisor in the given sample period. In addition, hotels begun to respond to reviews at different times (see Figure 1). We exploit this variation in our identification strategy. About 73% of the hotels are chain-affiliated, out of which 78% respond to reviews on Tripadvisor. 51% of the hotels

[4] Because of this our linking procedure is error-free.

are low-tier, 36% mid-tier, and 13% are high-tier.[5] 85% of high- and mid-tier hotels respond to reviews, and 64% of low-tier hotels respond to reviews.

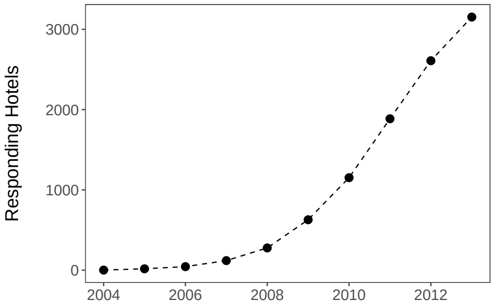


Figure 1: Cumulative Number of Hotels that Respond to Reviews by Year

Table 1: Expedia Rating Summary Statistics

	N	Mean	Std. Dev
Overall	1,855,527	4.01	1.1
Before Responding	895,810	3.94	1.16
After Responding	959,717	4.08	1.04

Table 1 presents the summary statistics of our main dependent variable: the star rating of an Expedia review. The mean star rating is 4.0, with a standard deviation of 1.1 and a median of 4.0. Furthermore, the average rating increased from 3.9 to 4.1 after the first management response, suggesting an increase in the average quality of hotels post-response.

Low- and high-rated hotels. To differentiate between low- and high-quality hotels in the pre-response period, we further classify hotels based on their Expedia ratings prior to their first response. We define a low-rated (high-rated) hotel as one whose average monthly rating has been mostly below (above) the median monthly rating for a given city and class (price tier) in the period before the hotel's first response. Specifically, we first compute every hotel's monthly average rating and median rating across all hotels in a given price tier and city in the pre-response period. Then, for every month, we identify hotels that are below (low-rated) and greater than or equal to (high-rated) the monthly median rating. Finally,

[5] We use the hotel class information provided by Tripadvisor to classify hotels.

we split the sample of hotels into two sub-samples based on the modal group (high-rated or low-rated) in all the months in the pre-treatment period. 55% of the hotels are classified as high-rated, and 80% of them respond to reviews during our sample period, whereas 67% of low-rated hotels respond to reviews.

Table 2: Average Rating of Low- and High-rated Hotels

	Low-rated	High-rated
Before Responding	3.6	4.2
	(1.2)	(1.0)
After Responding	3.9	4.2
	(1.1)	(1.0)

Note: Standard deviations in parentheses

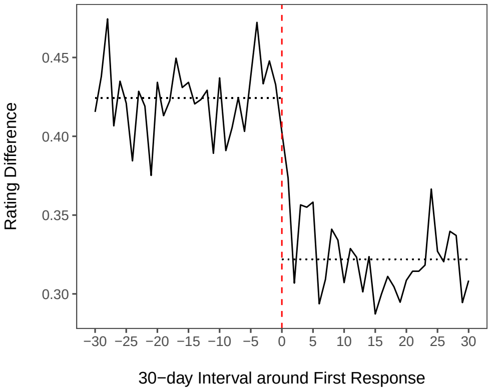


Figure 2: Difference between the Average Rating of High- and Low-rated Hotels

From Table 2, we can see that the difference between the average rating of low- and high-rated hotels decreased by 0.3 stars post-response, suggesting an improvement in the quality of low-rated hotels. In Figure 2, we plot the difference between the average rating of low- and high-rated hotels over 30 months before and after the first response.[6] We observe that the decline in the rating gap between the two types of hotels coincides with the time around the first response, indicating that the improvement in the ratings is not a result of a general time trend.

[6] The plot is generated using the 3,110 hotels that respond during our sample period.

3.4 Mining Latent Quality Issues from Consumer Reviews

In addition to using star ratings to measure changes in quality, we leverage large-scale review text data and a novel machine learning approach to mine the latent quality issues consumers raise in their reviews. For every hotel, we (i) extract entities frequently discussed by the reviewers in the pre-response period, (ii) assign a numerical sentiment score to each of these entities, and (iii) identify entities with negative sentiment.

Table 3: Examples of Entities and Sentiment Extracted from Expedia Reviews

Review Text	**Entity (Sentiment)**
The shower curtain was broken and the soap dispenser was empty and the night clerk was surly. That said, the day staff was friendly and helpful and they did try and fix the shower curtain. It was a negative experience, but the staff did their best to make up for it.	shower curtain (-.9), soap dispenser (-.9), night clerk (-.8), staff (.8)
Clean but outdated- our room had a weird smell like it used to be a smoking room. Comfortable beds. Great location!	room (-.3), smell (-.8), smoking room (0), beds (.9), location (.9)

We use Google Cloud Platform's state-of-the-art machine learning tool for NLP tasks to perform our analysis. The Cloud Natural Language API enables us to extract all the entities, e.g., "breakfast", "shuttle", mentioned in a text along with the sentiment score for each entity.[7]. Using this novel application, we extract each entity mentioned in a review and the sentiment associated with it. It is worth noting that the sentiment captured by this tool significantly differs from traditional sentiment analysis tools that give an overall sentiment score associated with each review or sentence. Furthermore, because the tool is pre-trained on a large and general dataset based on online user-generated content, it does not require

[7] For more information about this tool, see `https://cloud.google.com/natural-language/docs/analyzing-entity-sentiment`

thousands of correctly labeled samples that could be very costly to obtain. [8] In Table 3, we present the information obtained from this tool for two Expedia reviews.

Using this technique, we can efficiently scale up our analysis to millions of reviews and identify the problematic issues for each hotel in the pre-treatment period.[9] Specifically, we first obtain all the entities from each Expedia review with less than a perfect 5-star rating. Second, we find the top-5 most frequently mentioned entities in the reviews for each hotel before they start responding to reviews on TripAdvisor. Finally, we extract the sentiment score (which is a numerical value between -0.9 and 0.9) of the frequently mentioned entities from all the Expedia reviews that mention them.

Table 4: Statistics for the Top-5 Entities with Negative Sentiment

	Hotels	Unique Entities	Mean Sentiment		Negative Mentions	
			Before	After	Before	After
1st	414	39	-.135	.079	.555	.343
			(.419)	(.502)	(.221)	(.256)
2nd	425	41	-.135	.030	.553	.351
			(.428)	(.458)	(.228)	(.256)
3rd	516	53	-.117	.031	.534	.376
			(.405)	(.482)	(.231)	(.291)
4th	579	67	-.127	.010	.541	.369
			(.397)	(.453)	(.228)	(.290)
5th	624	81	-.144	.008	.545	.378
			(.405)	(.430)	(.228)	(.299)

Note: Standard deviations in parentheses

Because our primary focus of this analysis is to see whether hotels resolve the issues that reviewers mention, we limit our hotel sample to hotels for which the X most frequently mentioned entity ($X \in \{1^{st}, 2^{nd}, 3^{rd}, 4^{th}, 5^{th}\}$) has a negative average sentiment score in the pre-response period. In doing so, we are left with 1,667 distinct hotels out of 4,180 hotels in

[8] The results presented in this paper are likely not algorithm-specific. To support this statement, we verified the sentiment output of the Google Cloud Natural Language API with that of a popular package for aspect-based sentiment mining developed by ScalaConsultants from `https://github.com/ScalaConsultants/Aspect-Based-Sentiment-Analysis` and we find that the two algorithms have an average agreement of 93%.

[9] The number of reviews is lower in this analysis because about 25% have no text.

our dataset. The number of hotels, unique entities, mean sentiment score, and the fraction of reviews with negative mentions (i.e., the fraction of reviews where the entity was mentioned and had a negative sentiment score) for each of the top five entities during the pre-and post-response periods are reported in Table 4. The high value of unique entities indicates that the issues consumers complain about are very diverse across hotels. We observe that the mean sentiment for these entities increases significantly and even turns positive in the post-treatment period. Further, the fraction of reviews with a negative-sentiment entity also diminishes significantly in the post-treatment period for each of the top five entities. These descriptive results suggest that hotels improve on frequently mentioned issues. In Section 5, we present a formal analysis confirming these descriptive results.

4 The Effect of Paying Attention to Reviews on Ratings

4.1 Empirical Strategy

Our objective is to investigate whether online reviews help hotels improve their quality. There are several challenges to identification that we need to address. First, whether hotels pay attention to reviews is something we cannot directly observe. Therefore, we need a way to identify which hotels are more likely to read and pay attention to reviews. We do so by using management response on Tripadvisor as a proxy for whether a hotel pays attention to the reviews written by its customers. In doing so, we are making two assumptions. The first one is that hotels that respond to reviews are more likely to pay attention to reviewers than those that do not respond. This assumption is supported by past literature such as Chevalier et al. (2018) and Proserpio and Zervas (2017). The second assumption is that when hotels begin to listen or pay more attention to reviews, they also start responding to them.[10] It is worth noting that for our treatment to be valid, we do not require non-responding hotels not to pay attention; we simply require a change in the behavior of hotels that begin to respond to reviews. While both these assumptions are imperfect, the fact that in Figure 3, we observe changes in ratings around the time of the first response suggests that our assumptions are not far from what really happens.

Second, hotels *decide* when and to which reviews to respond. Therefore, we estimate the average treatment effect on the treated (ATT); in other words, the impact of listening to consumer reviews for hotels that have decided to listen to them. The ATT is defined as:

$$E[Y_{i1} - Y_{i0}|D = 1], \tag{1}$$

where Y_{i1} is the rating for hotel i from the treatment condition, Y_{i0} is the rating from the control condition, and $D = 1$ indicates that hotel i is among those that are treated (i.e., among those that post management responses). Because we are not using experimental data in our analysis, the counterfactual ratings for hotel i, i.e., Y_{i0} are not observed. This implies that we need to construct a credible control group for the hotels that decide to respond to reviews. A good starting point would be to use ratings from the pre-treatment period (i.e., rating of hotel i before it started responding to reviews) as a control. This is problematic because the timing of the first response is likely endogenous. Thus this strategy could lead to biased results. Instead, we exploit the fact that, during our sample period, (i) not all hotels respond to reviews, and (ii) hotels begin responding to reviews at different times (see Figure 1). In doing so, we implement a generalized difference-in-differences (DD) identification strategy that compares changes in ratings of hotels whose responding status has changed with those of hotels whose status has not (or not yet) changed (Goodman-Bacon 2021).

[10] In simple terms, the identification of the treatment relies on the existence of a correlation (and not a causal relationship) between responding and paying attention (or paying more attention) to reviews. To some extent, this is a situation similar to the instrumental variables strategy where the instrument does not need to cause changes in the endogenous independent variable, but a simple correlation suffices (Cameron and Trivedi 2005).

Third, estimating our DD on the Tripadvisor sample will likely generate biased results. This is because, as discussed in several papers (Proserpio and Zervas 2017, Proserpio et al. 2021), the mere presence of management responses affects consumers' reviewing behavior which, in turn, can affect the average ratings of the hotel. To address this issue, we exploit the fact that hotels respond to reviews on Tripadvisor, but they rarely do so on Expedia, and estimate our DD strategy using Expedia ratings as the dependent variable and identify responding hotels using Tripadvisor responses. Because there are no responses on Expedia, this allows us to effectively eliminate the effect of management responses on ratings.[11]

Finally, any change in ratings we observe may be due to factors other than the content of the reviews. We directly address this concern in Section 5, where we use NLP algorithms and the same identification strategy to provide evidence that hotels do improve on aspects that are frequently mentioned in the reviews they receive. In addition, in Section 7, we present a battery of robustness checks aimed at ruling out several alternative explanations for our results.

The leading approach in applied work that utilizes DD with variation in treatment timing has been to estimate a regression with cross-sectional units (α_i) and time periods (τ_t), and a treatment dummy (D_{it})—commonly known as the two-way fixed effects (TWFE) model. The standard TWFE model compares changes in the outcome of units whose treatment status changes over time with those whose treatment status does not change, yielding a weighted average of all possible permutations of pairwise difference-in-differences estimators.

$$Y_{it} = \alpha_i + \tau_t + \beta D_{it} + \epsilon_{it}. \tag{2}$$

[11] An implicit assumption behind this strategy and that of similar papers (Chevalier et al. 2018, Proserpio and Zervas 2017, Wang and Chaudhry 2018) is that there are no or minimal cross-platform spillovers. In our case, this means that reviewers are unlikely to view responses on Tripadvisor before leaving a review on Expedia. While this assumption is not directly testable, travelers receive a direct link from Expedia after their stay that directs them to a page where they can write a review. Therefore, an average customer is unlikely to view Tripadvisor's responses at the time of posting a review on Expedia.

However, recent econometrics literature has raised concerns about the possibility of negative weights using the TWFE estimator when treatment timing is staggered and heterogeneity in treatment effects within units over time or between groups of units treated at different times (De Chaisemartin and d'Haultfoeuille 2020, Sun and Abraham 2020, Goodman-Bacon 2021). In such a case, the aggregated impact could be positive even if most individual pairwise DD estimators are negative.

To address this issue, we start by presenting the main results using the estimator proposed by Callaway and SantAnna (2020) (henceforth CS) and then replicate the results using the traditional TWFE estimator.[12] A few recently published papers (Ang 2021, Prager and Schmitt 2021) have also employed the CS estimator to resolve the issue.

Let $Y_{it}(0)$ be unit i's untreated potential outcome at time t and $Y_{it}(g)$ be unit i's potential outcome in time period t if they become treated in period g. We assume that there is no anticipation of treatment so that $Y_{it}(0) = Y_{it}(g)$ for all $t < g$. We also assume that once a unit is treated, it remains treated.[13] Then group-time average treatment effect be defined as $ATT(g,t) = \mathbb{E}[Y_{it}(g) - Y_{it}(0)|G_{ig} = 1] \; \forall \; t \geq g$, where G_{ig} is a binary indicator that equals 1 if hotel i starts responding in period g. $ATT(g,t)$ is the average effect of participating in the treatment for hotels in the group g at period t. The CS estimator aggregates all the group-time average treatment effects in a way that the weighted average does not suffer from the same drawbacks as the standard TWFE estimator. Further, if one is interested in estimating the average treatment effect by the length of exposure to treatment (e.g., to evaluate pre-treatment trends), one could estimate a weighted average of $ATT(g,t)$ conditional on $t - g = l$, where $l \in \mathbb{Z}$ is the desired length. Each group-time treatment effect, $ATT(g,t)$, is estimated by computing a weighted DD estimate where the reference time period is $g - 1$. The treatment group includes the observations with $G_{ig} = 1$, whereas

[12] We do not solely rely on the CS estimator because it does not allow for the inclusion of time-varying controls.

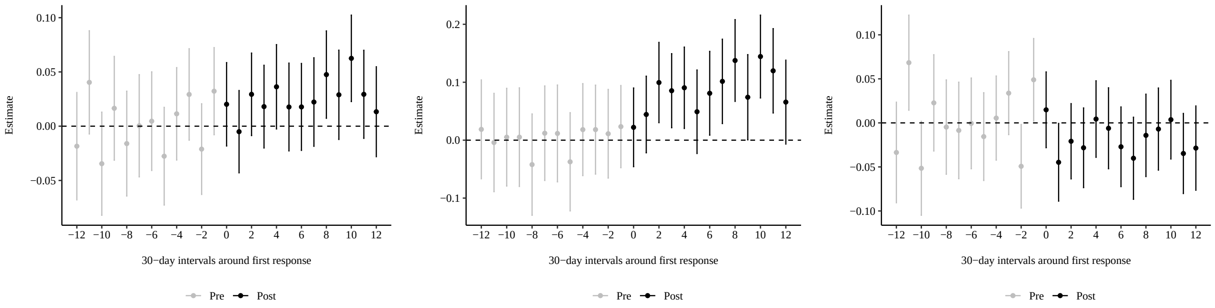
[13] Our data supports this assumption. For instance, of all the hotels that started responding before 2013 (the last year in our data) about 87%, 85%, 81% and 71% left a response in the last 12, 9, 6, and 3 months of 2013, respectively. In Section 7, we also consider the frequency of response as a continuous treatment variable, which may vary across the hotels.

the control group contains the observations with $G_{ig} = 0$ and have not yet been treated by time t. We refer to Callaway and SantAnna (2020) for the technical details on estimation and inference. We implement the estimator in R using the package DID.

4.2 Identification Check

The fundamental identifying assumption of our strategy that allows for a causal interpretation of our estimates is that Expedia ratings of treated and control hotels would have evolved similarly in the absence of treatment. Callaway and SantAnna (2020) show that the group-time treatment effects are identified from data on Y_{it}, and G_{ig} as long as the parallel trends assumption holds in addition to standard independence and support conditions. Formally, the parallel trends assumption for our setting can be stated as follows: for each g, h, and t such that $g \leq t \leq h$, $E[Y_{it}(0) - Y_{i,t-1}(0)|G_{ig} = 1] = E[Y_{it}(0) - Y_{i,t-1}(0)|G_{ih} = 1]$. The assumption requires that if treated groups had not been treated, their outcome would evolve in the same way as groups that have not yet been treated. While the parallel trends assumption is not testable because we do not observe counterfactual ratings for treated hotels, given our long observation period, we can test whether ratings for responding and non-responding hotels evolved similarly in the pre-treatment period. Specifically, the control group is set to the group of hotels that have not started responding during the sample period. This includes all never-treated hotels (those that do not start responding during our sample period) and additional units that eventually start responding but have not done so yet.

To compare the pre-treatment trends, we compute the treatment effect by the length of exposure (or event study) using the CS estimator. We define $\theta(l)$ as the weighted average of $ATT(g, t)$ for all t and g such that $t - g = l$, that is, $\theta(l) = \frac{1}{\kappa_l}\Sigma_g \omega_g ATT(g, g + l)$, where ω_g is proportional to the number of hotels with treatment month g, and κ_l normalizes the weights so that they sum up to one. The parameter $\theta(l)$ can be interpreted as an event study. More precisely, for $l < 0$, $\theta(l)$ captures the average effect of participating in the treatment for the group of hotels that are exactly l periods away from the period of the first response.



(a) All Hotels (b) Low-rated Hotels (c) High-rated Hotels

Figure 3: Event Study Plots: The evolution of treatment effect as a function of a hotel's decision to begin responding to reviews. The solid dots plot the estimates, and the vertical bars represent their respective 95% confidence intervals.

Conversely, for $l > 0$, $\theta(l)$ captures the average effect of participating in the treatment for the group of hotels that have responded for exactly l periods.

As discussed previously, we hypothesize that because low-rated hotels have more room for improvement, the effect of listening to reviews should be stronger for low-rated hotels compared to high-rated hotels. Therefore, we report the results of the event study for the three groups of hotels: all, low-rated, and high-rated (as defined in Section 3).

Figure 3 reports our estimates (and 95% confidence intervals) of $\theta(l)$ for the overall sample, the sample with only low-rated hotels, and the sample with high-rated hotels, for l running between -12 and 12 months (i.e., one year before and after the treatment starts.).

A few things emerge from this figure. First, we can see that before hotels begin responding to reviews (Interval 0), ratings evolve in a similar way (estimates are indistinguishable from zero), particularly for low-rated hotels. Second, while post-treatment estimates are noisy and close to zero for all high-rated hotels, there is a visible and consistent upward trend for low-rated hotels (see Figure 3b). This finding supports our hypothesis that these are the hotels that can improve their ratings the most.

4.3 Results

We now estimate the aggregate effect of responding to reviews on Tripadvisor on Expedia ratings. We first estimate the impact using the CS estimator, aggregating the data at the hotel-year-month level. We report the overall treatment effect, $\bar{\theta}$, which is a weighted average of $ATT(g,t)$ for $t \geq g$ over all groups and time periods, that is, $\bar{\theta} = \frac{1}{\kappa}\Sigma_g \Sigma_{t>g} \omega_g ATT(g,t)$, where again ω_g is proportional to the number of hotels with treatment month g, and κ normalizes the weights so they sum up to one.

We present the treatment effect estimates for the sample, including (i) all the hotels, (ii) only low-rated hotels, and (iii) only high-rated hotels in Table 5. The estimate reported in the first column of Table 5 suggests that after hotels started responding on Tripadvisor, their rating increased by 0.04 stars on average. However, all of the overall positive effects are driven by the sample of initially low-rated hotels, whose ratings increase by 0.08 stars on average (see Column 2). In contrast, the effect is relatively small and statistically insignificant for high-rated hotels in the post-treatment period (see Column 3).[14]

Table 5: Effect on Ratings (CS)

	All	Low-rated	High-rated
After Responding	0.039^{**} (0.018)	0.115^{***} (0.031)	-0.025 (0.020)
Observations	220050	92323	127727

Significance levels: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.
Note: The dependent variable is the average rating of hotel j at time t. Cluster-robust standard errors (at the individual hotel level) are shown in parentheses.

Because the CS estimator does not allow for the inclusion of time-varying covariates, we next report the results using the standard TWFE model. Specifically, we estimate the following regression model at the review level.

$$Stars_{ijt}^{E} = \beta \text{After Responding}_{ijt}^{TA} + \mathrm{X}_{jt}^{'}\gamma + \alpha_j + \tau_t + \epsilon_{ijt}, \tag{3}$$

[14] The insignificant effect for high-rated hotels is likely due to the fact that while listening to reviews helps low-rated hotels improve their ratings, it helps high-rated hotels maintain high ratings.

where the dependent variable, $Stars^E_{ijt}$, is the Expedia rating of review i written by hotel j at year-month t. After Responding$^{TA}_{ijt}$, whose coefficient is of interest, is a binary indicator that equals 1 if review i was written on Expedia after hotel j started responding to reviews on Tripadvisor at year-month t.; it captures the effect of responding to reviews on Tripadvisor on Expedia ratings. We include hotel fixed effects, α_j, to account for time-invariant hotel characteristics that can affect ratings, and year-month fixed effects, τ_t, to account for time-varying shocks to ratings common to all hotels. In addition, we control for time-varying review environment characteristics, X_{jt} which include the cumulative average rating on Expedia, and the logarithm of the cumulative number of reviews of hotel i on Expedia, both taken at time $t - 1$. These covariates help us control for the effect that past reviews have on future ratings. Finally, to account for serial correlation, we cluster standard errors at the hotel level. Like before, we estimate Equation 3 separately for all the hotels, low-rated hotels, and high-rated hotels.

Table 6: Effect on Ratings (TWFE)

	All	Low-rated	High-rated
After Responding	0.047***	0.110***	0.008
	(0.011)	(0.019)	(0.012)
Observations	1855467	748313	1107152
R^2	0.206	0.205	0.146

Significance levels: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.
Note: The dependent variable is the rating of review i of hotel j at time t. Cluster-robust standard errors (at the individual hotel level) are shown in parentheses. All specifications include hotel and year-month fixed effects.

We start by reporting the results without any controls for the sample, including all the hotels in Column 1 of Table 6. The coefficient of interest, After Responding$^{TA}_{ijt}$, is positive and statistically significant, suggesting that once hotels start responding on Tripadvisor, their Expedia rating increases by 0.047 stars on average. Columns 2 and 3 report the estimates for low- and high-rated hotels, respectively. Again, we find that, while there is no effect (the coefficient estimate is small and statistically insignificant) for high-rated hotels, the estimate

Table 7: Effect on Ratings with Covariates (TWFE)

	All	Low-rated	High-rated
After Responding	0.044***	0.098***	0.010
	(0.010)	(0.017)	(0.011)
Avg. Expedia Rating$_{t-1}$	0.276***	0.272***	0.266***
	(0.021)	(0.033)	(0.024)
Log Number of Expedia Reviews$_{t-1}$	-0.002	-0.004	0.001
	(0.005)	(0.008)	(0.006)
Observations	1851316	746464	1104851
R^2	0.207	0.206	0.147

Significance levels: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.
Note: The dependent variable is the rating of review i of hotel j at time t. Cluster-robust standard errors (at the individual hotel level) are shown in parentheses. All specifications include hotel and year-month fixed effects.

is positive ($\beta = 0.1, p < 0.01$) and statistically significant for low-rated hotels.[15] Table 7 reports the results including review environment controls. The estimate remains positive, statistically significant, and similar in magnitude to the previous table.

To put the main estimate into context, we compare it with the average ratings of low-rated hotels in the pre-treatment period, which is 3.6 stars (SD=1.2 stars). Considering our preferred specification (Column 2 of Table 7), our estimate suggests that, on average, low-rated hotels improve their rating by about 2.5%.

Overall, the results presented in this section suggest that low-rated hotels primarily drive the effect of listening to reviews. In Online Appendix A, we show that larger firms like chain or upscale hotels gain the most from online customer reviews, likely because they have more resources to invest in monitoring reviews and making improvements in quality.

The implications of these results are important as they demonstrate that online reviews are a valuable source of information that low-rated hotels can use to make quality improvements which, in turn, reduce the quality (rating) gap between low- and high-rated hotels.

[15] The CS and the TWFE estimators yield similar results, suggesting that the negative weights problem is a relatively minor issue in our context. Indeed, on diagnosis, we find that only about 17% of the weights are negative, summing up to merely -.098, while the positive weights sum up to 1.098.

5 Linking Rating Improvements with the Content of the Reviews

So far, we have used numerical ratings of reviews to measure changes in quality that hotels experience after listening to their customers via online reviews. While this quantitative data is observable and readily available, it does not provide us with the level of detail observed in the raw unstructured text data. Therefore, we leverage the large-scale textual data from the reviews and responses and use novel ML-based NLP techniques to provide further evidence that hotels raise their quality (ratings) by listening to their customers' feedback online.

Specifically, we employ the same strategy discussed in Section 4, and measure changes in the sentiment of problematic entities in the post-treatment period. This analysis enables us to investigate problems unique to each hotel that may not be reflected in the overall star ratings of the reviews. As an illustration of our approach, consider the following scenario: a hotel has had problems with the swimming pool and the breakfast buffet before January 2012. Guests of this hotel have noticed and complained about the pool and the breakfast buffet in their reviews. The hotel management begins reading the reviews and responding to them in January 2012, assuring the customers that the problems with the pool and the breakfast buffet were being addressed. If this is the case, we should see that the complaints in the reviews about the pool and the breakfast fade over time.

As discussed in Section 3.4, to implement this approach, we use NLP tools to parse reviews and extract the most frequently mentioned issues in the pre-treatment period for each hotel. We then check whether these issues are resolved in the post-treatment period. In the above scenario, the entities "pool" and "breakfast" would have, on average, a negative score across reviews. If the hotel had truly paid attention to the pool and breakfast issues brought up by its customers and acted on the complaints, we would expect the sentiment associated with entities pool and breakfast to increase after January 2012, from negative to neutral or positive.

5.1 Main Results

Having provided the descriptive evidence in Section 3.4 that hotels pay attention to the issues mentioned in the reviews and take action on them, we next move to a more formal analysis. To formalize this analysis in a regression framework, we need to make an important decision related to the reviews not mentioning the focal entity. Mentions of entities, even those most frequently mentioned, are very sparse and most reviews do not mention them.[16] However, non-mentions can carry important information because they could suggest that an entity that was an issue in the past is no longer one. Therefore, to account for whether an entity is mentioned or not, we assign a neutral score of 0 to every review that does not mention the focal entity.[17]

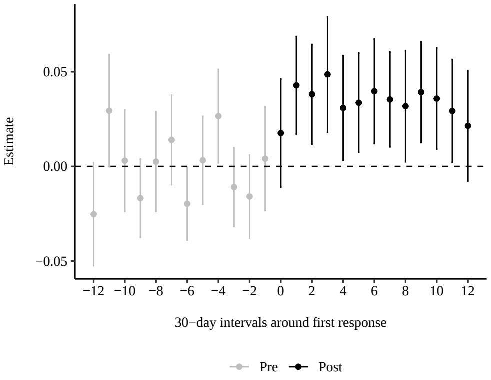


Figure 4: Event Study Plot: The evolution of treatment effect on the sentiment of the most frequently mentioned problematic entity as a function of a hotel's decision to begin responding to reviews. The solid dots plot the estimates, and the vertical bars represent their respective 95% confidence intervals.

Similarly to what we did for the star ratings, we first compute the treatment effect by the length of exposure using the CS estimator. For brevity, we focus on the top-mentioned entity. Figure 4 reports our estimates (and 95% confidence intervals) of $\theta(l)$ for l running between -12 and 12 months (i.e., one year before and after the treatment starts.). In the pre-

[16] For example, 82% of the reviews do not mention the first-most frequently mentioned entity; and 94% of the reviews do not mention the fifth-most frequently mentioned entity.

[17] In Table 23 in Online Appendix B, we show that our results are not sensitive to the exclusion of reviews not mentioning the entities.

treatment period, the estimates are close to zero and statistically insignificant, while they become positive and significant in the post-treatment period, suggesting that the sentiment of the most mentioned entity improves after hotels begin to respond to reviews.

We next present the estimates using the CS estimator for each one of the top five most mentioned entities in Table 8. The estimates are positive and significant for four out of the five entities. These results suggest that there is a net positive change in the average sentiment of the frequently mentioned entities for the hotels in our sample. Moreover, the effect size decreases as we move from the 1st to the 5th most frequently mentioned entity suggesting that hotels put more effort into addressing issues that reviewers mention more often.

Table 8: Effect on the Sentiment of Entities (CS)

	1st Most	2nd Most	3rd Most	4th Most	5th Most
After Responding	0.037^{**} (0.012)	0.034^{**} (0.009)	0.025^{**} (0.008)	0.022^{**} (0.006)	0.021^{**} (0.007)
Observations	18478	20481	26581	31081	35076

Significance levels: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.
Note: The dependent variable is the average sentiment score of a top-5 entity of hotel j at time t. Cluster-robust standard errors (at the individual hotel level) are shown in parentheses.

We then turn to the TWFE estimator and estimate Equation 3 using the sentiment score of each one of the top-5 entities as our dependent variable, without and with time-varying control variables. We report these results in Tables 9 and 10. In line with the CS estimates, we find that the coefficient of interest is positive and statistically significant for each of the top five entities.

5.2 Additional Analyses

We complement and validate the above findings with additional analyses in which we show that (i) the sentiment of entities of hotels with more homogeneous complaints improves more; (ii) the sentiment of entities part of reviews with a response improves more; and (iii) the sentiment of entities classified as non-issues do not change.

Table 9: Effect on the Sentiment of Entities (TWFE)

	1st Most	2nd Most	3rd Most	4th Most	5th Most
After Responding	0.0179***	0.0136***	0.0088***	0.0100***	0.0078***
	(0.0027)	(0.0019)	(0.0013)	(0.0011)	(0.0010)
Observations	92275	104597	168282	180442	218077
R^2	0.0496	0.0160	0.0165	0.0113	0.0135

Significance levels: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.
Note: The dependent variable is the sentiment score of a top-5 entity in review i of hotel j at time t. Cluster-robust standard errors (at the individual hotel level) are shown in parentheses. All specifications include hotel and year-month fixed effects.

Table 10: Effect on the Sentiment of Entities with Covariates (TWFE)

	1st Most	2nd Most	3rd Most	4th Most	5th Most
After Responding	0.0170***	0.0136***	0.0087***	0.0096***	0.0076***
	(0.0026)	(0.0019)	(0.0013)	(0.0011)	(0.0010)
Avg. Expedia Rating$_{t-1}$	0.0077*	0.0007	0.0088***	0.0031*	0.0035**
	(0.0042)	(0.0034)	(0.0027)	(0.0020)	(0.0020)
Log Number of Expedia Reviews$_{t-1}$	0.0046**	-0.0005	0.0001	0.0008	0.0002
	(0.0018)	(0.0014)	(0.0010)	(0.0008)	(0.0007)
Observations	91938	104244	167853	179953	217555
R^2	0.0501	0.0162	0.0168	0.0113	0.0137

Significance levels: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.
Note: The dependent variable is the sentiment score of a top-5 entity in review i of hotel j at time t. Cluster-robust standard errors (at the individual hotel level) are shown in parentheses. All specifications include hotel and year-month fixed effects.

Homogeneous vs heterogeneous issues. To this end, for each hotel, we compute the fraction of reviews where the focal entity is mentioned with negative sentiment in the pre-treatment period. We call this variable *Pre NegMentions*. This variable captures the level of homogeneity in complaints across reviewers. Our hypothesis is that the effect on the entities' sentiment should be stronger for those hotels that receive a more homogeneous set of complaints than those whose reviews are a more heterogeneous set of complaints. This is because more concentrated reviews tell something about the severity of the issue, leading to a higher expected benefit for the hotel. We add the interaction term After responding × Pre NegMentions to Equation 3, and re-estimate it using the sentiment score of each one of the top five entities as our dependent variable. The coefficient of the interaction term

Table 11: Heterogeneous Effects on the Sentiment of Entities

	1st Most	2nd Most	3rd Most	4th Most	5th Most
After Responding	-0.0108***	-0.0013	0.0006	0.0004	0.0003
	(0.0029)	(0.0044)	(0.0018)	(0.0016)	(0.0017)
After Responding × Pre NegMentions	0.4852***	0.3131***	0.2203***	0.2880***	0.2561***
	(0.0388)	(0.0962)	(0.0510)	(0.0423)	(0.0571)
Avg. Expedia Rating$_{t-1}$	0.0085**	0.0011	0.0102***	0.0035*	0.0046**
	(0.0040)	(0.0033)	(0.0028)	(0.0020)	(0.0021)
Log Number of Expedia Reviews$_{t-1}$	0.0021	-0.0015	-0.0004	0.0005	0.0001
	(0.0017)	(0.0014)	(0.0010)	(0.0007)	(0.0006)
Observations	91938	104244	167853	179953	217555
R^2	0.0517	0.0167	0.0170	0.0117	0.0140

Significance levels: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.
Note: The dependent variable is sentiment score of a top-5 entity in review i of hotel j at time t. Cluster-robust standard errors (at the individual hotel level) are shown in parentheses. All specifications include hotel and year-month fixed effects.

captures the differential improvement in the entity sentiment for hotels that received a more homogeneous set of complaints in the pre-treatment period.[18]

We report these results in Table 11. The coefficient of interest is positive and statistically significant for each one of the entities, implying that the overall improvement in the sentiment of problematic entities is stronger for hotels for which the entity has been mentioned more often in the pre-treatment period.

Effect for entities in responded reviews. In addition, we analyze how the effect on sentiment varies between problematic entities mentioned in the reviews with and without a management response. Our hypothesis is that the sentiment improvement should be higher for problematic entities (i.e., average pre-response sentiment is negative) mentioned in the reviews that received a response; this is because the hotel likely paid more attention to such reviews since they responded to them. To test this hypothesis, for every top-5 frequently

[18] A concern related to this analysis is that the fraction of mentions of the entity is likely correlated with its sentiment in the pre-treatment period (it is likely that the more problematic an issue is, the more is discussed by different reviewers in the same negative context), which could also drive a hotel to react more. Therefore, as a robustness check, we include an additional interaction in our model, After responding × Pre Sentiment, where the variable Pre Sentiment is the average sentiment of the entity in the pre-treatment period. We report these results in Table 24 in Online Appendix B.

Table 12: The Moderating Effect of Receiving a Response on the Sentiment of Entities

	1st Most	2nd Most	3rd Most	4th Most	5th Most
After Responding	0.009***	0.010***	0.006***	0.008***	0.006***
	(0.003)	(0.002)	(0.002)	(0.001)	(0.001)
After Responding × Responded	0.091**	0.065**	0.059	0.019	0.071
	(0.043)	(0.029)	(0.045)	(0.020)	(0.043)
Avg. Expedia Rating$_{t-1}$	0.005	-0.000	0.011***	0.002	0.004**
	(0.004)	(0.004)	(0.003)	(0.002)	(0.002)
Log Number of Expedia Reviews$_{t-1}$	0.005**	-0.000	0.000	0.001	0.000
	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)
Observations	89548	95115	135110	166522	205422
R^2	0.064	0.022	0.025	0.012	0.019

Significance levels: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.
Note: The dependent variable is the sentiment score of a top-5 entity in review i of hotel j at time t. Cluster-robust standard errors (at the individual hotel level) are shown in parentheses. All specifications include hotel and year-month fixed effects.

mentioned entity we create a binary indicator, *Responded*, that equals 1 if it was mentioned in a review that received a response, otherwise 0. We then add the interaction term After responding × Responded to Equation 3, and re-estimate it using the sentiment score of each one of the top five entities as our dependent variable. Consistent with our expectation, we find that the estimate of the interaction term's coefficient is positive, and statistically significant for the two most frequently mentioned entities (see Table 12).

Effect for non-issues. Finally, we verify that we are actually capturing changes in the quality of entities that were problematic, e.g., an overall increase in sentiment across entities, we repeat the analysis discussed above for entities that are not issues, i.e., with positive sentiment in the pre-response period. The intuition behind this test is that if an entity is not a problem for the hotel, we should not observe changes of its sentiment in the post-treatment period. In Figure 5, we compare the estimates for these entities with the ones for entities with a negative sentiment (reported in Table 10). As expected, we observe a small and indistinguishable from zero effect for the sentiment of entities with positive average sentiment in the pre-response period.

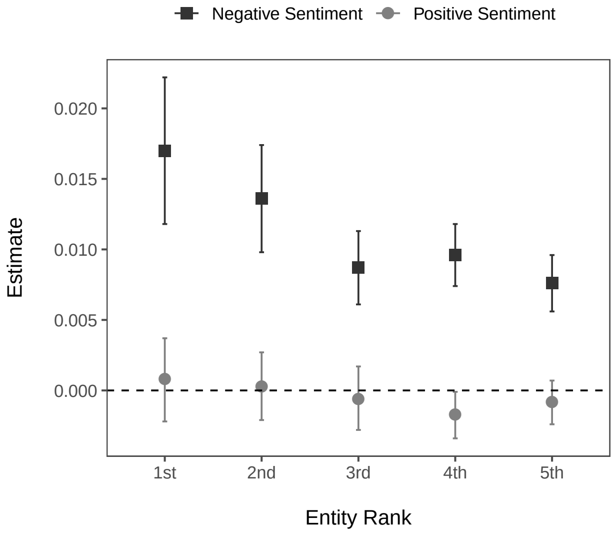


Figure 5: Comparison of Estimates for Entities with Average Negative and Positive Sentiment in the Pre-response Period

Overall, the results discussed in this section are consistent with and provide further validity to our main findings discussed in Section 4. They demonstrate the presence of a robust online environment in which customers are empowered to alert the hotels about quality issues which, in turn, enables hotels to address these issues offline.

6 The Moderating Impact of Canned Responses

In Section 5, we focused on the demand side and identified latent quality issues of each hotel by mining consumer reviews. Here we focus on the supply side and investigate whether the way in which hotels respond to reviewers on the platforms reveals how seriously hotels take their customers' feedback. Intuitively, hotels that copy and paste the same response to different reviews, or that write canned responses, likely care less about and pay less attention to what their customers have to say. Because of this, we should observe smaller or no change in the ratings of these hotels after they begin to respond to reviews.

To test this hypothesis, we leverage a deep learning method to analyze the content of management responses and create a measure of *response similarity* for each hotel and calendar year. High response similarity indicates that a hotel is responding in the same way to every review, potentially suggesting that it has little interest in what its customers have to say. Low response similarity indicates that the hotel customizes its reply to each review, potentially suggesting that it reads and cares about what its customers have to say.

To create a measure of response similarity, we employ the Universal Sentence Encoder (USE) (Cer et al. 2018) that encodes a sentence or a paragraph to a numerical vector of a specified dimension. USE is similar to (and inspired by) representation learning techniques such as *word2vec* (Mikolov et al. 2013) that transform words to multidimensional vectors. The basic intuition behind word embedding techniques is that words that have contextual similarities are mapped to nearby points. For example, one can use vector algebra to show that "king" - "man" + "woman" will be very similar to the vector representation of the word "queen". USE extends this idea to sentences so that similar documents or paragraphs that are similar to each other will have similar vector representations. For instance, "How old are you?" and "What is your age?" will have very similar representations with a high cosine similarity score, despite having no words in common. USE is optimized and trained to analyze sentences and short paragraphs and is available as a free library in Google's TensorFlow Hub.[19]

For every hotel and year in our dataset, we consider the response to a review as a document and find its vector representation, and then compute the pair-wise cosine similarity between all possible pairs of responses. In Table 13, we report three pairs of responses with varying degrees of similarity. One can observe that USE captures the similarity between sentences very well.

We use the pair-wise similarity scores to compute the number of unique responses each hotel management writes in a year. Since USE provides us with a continuous measure of

[19] For more information about this tool, see: `https://tfhub.dev/google/universal-sentence-encoder-large/3`. USE uses two kinds of encoders. The first model uses transformer architecture by Vaswani et al. (2017) to develop encodings of sentences by calculating context-aware representations of words in sentences. The second model creates sentence embeddings by averaging values of embeddings from words and bi-grams, which are then fed into a feed-forward deep neural network. We use the first type of encoder because of its slightly higher accuracy.

Table 13: Cosine Similarity Between Responses Based on USE

Response 1	**Response 2**	**Similarity**
Thank you for telling us about your stay. I would like to discuss your stay further with you if possible to get a better understanding of your concerns. Please contact me at your convenience.	*Thank you for sharing your comments with us. I would like to discuss in further detail with you about your stay. Please contact me at your convenience. Thank you.*	0.91
Thank you for telling us about your stay. I would like to discuss your stay further with you if possible to get a better understanding of your concerns. Please contact me at your convenience.	*Thank you for telling us about your stay. You will be happy to know that our sinks are all standard size throughout all Hampton hotels as well as the towel shelf centrally located above the toilet. We hope to see you again upon your next trip to Wichita.*	0.45

similarity, we first need a way to identify and count unique responses. We consider two responses to be the same if the cosine similarity between them is above a threshold which we denote by η. Because we want responses to be very similar, we use a very high threshold, i.e. $\eta = 0.9$.[20] We call a response *unique* if the cosine similarity with all other responses by the hotel is below η. Let N and N_u denote the total number of responses and the number of unique responses for a hotel-year, respectively. Then, $1 - \frac{N_u}{N}$ gives us the measure of response similarity for each hotel-year. We denote this measure by θ. Therefore, θ can be interpreted as the percentage of non-unique responses. For example, $\theta = 0.25$ implies that one out of four responses is not unique, whereas $\theta = 0.75$ implies that three out of four responses are not unique. Thus, a hotel with a high θ uses more canned responses to address its reviews, while a hotel with a lower θ gives a more personalized response to each review.

Since the distribution of the similarity scores across hotel-years is extremely right-skewed (only ∼1% of scores lie above 0.7), to estimate the impact of response similarity, we create a binary indicator, Similarity_{jt}, that equals 1 if θ is greater than 0.8, 0.7, or 0.6, else equals 0.[21]

[20] We obtain similar results using $\eta = 0.95$.

[21] While we test the sensitivity of our results to different thresholds, we choose relatively high thresholds to select hotels that use canned responses for most of the reviews.

Using the subsample of hotels we identified as low-rated in Section 3, we estimate Equation 3 with the addition of the interaction $\text{After}^{TA}_{jt} \times \text{Similarity}_{jt}$ which captures the moderating effect of canned responses on quality improvements, if any.

Table 14: Moderating Effect of Response Similarity on Ratings ($\eta = 0.90$)

	$\theta > .8$	$\theta > .7$	$\theta > .6$
After Responding	0.098***	0.099***	0.100***
	(0.017)	(0.017)	(0.017)
After Responding × Similarity	-0.104***	-0.071**	-0.079*
	(0.029)	(0.034)	(0.042)
Avg. Expedia Rating$_{t-1}$	0.269***	0.270***	0.269***
	(0.032)	(0.033)	(0.033)
Log Number of Expedia Reviews$_{t-1}$	-0.004	-0.004	-0.004
	(0.008)	(0.008)	(0.008)
Observations	746464	746464	746464
R^2	0.206	0.206	0.206

Significance levels: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.
Note: The dependent variable is the rating of review i of hotel j at time t. Cluster-robust standard errors (at the individual hotel level) are shown in parentheses. All specifications include hotel and year-month fixed effects.

We report the results for the three different thresholds in Table 14. In line with our hypothesis, we find that the coefficient of the interaction is negative for all three thresholds and statistically significant, suggesting that the main effect of listening to reviews is lower for hotels that have a high response similarity.

The analysis discussed in this section suggests that the way hotels respond to customers carries important information about whether they care and act on issues raised by their customers or simply ignore them. Hotels that tend to use canned replies are less likely to improve in quality because they are not invested in listening to their customers, while those that care about addressing their customers' problems are also more likely to listen to their consumers and, in turn, more likely to obtain higher gains in quality.

7 Robustness Checks

The identifying assumption of our strategy is that there are no unobservables correlated with both hotel ratings and the timing of the first response. While the event study plot we presented in Figure 3 reduces the concerns regarding such unobservables, the possibility of the existence of a confounder still remains. In what follows, we proceed to describe several tests in support of our main findings. Here we present the results using star-ratings of low-rated hotels as the dependent variable and in Online Appendix C using the sentiment of the most frequently mentioned problematic entity.

Differences between responding and non-responding hotels. One may argue that non-responding hotels are *different* from responding hotels, and these differences are driving the results. To reduce this concern, first, we include in Equation 3, hotel-specific time trends which allow for time-variant (according to a pre-specified function, which is linear in our case) confounders correlated with the treatment status. We report these results in column 1 of Tables 15. We find that the coefficient of interest is positive, statistically significant, and similar to our main estimate reported in the third column of Table 7. Second, we exclude non-responding hotels from the estimation, thus limiting the sample to hotels that eventually respond. We report these results in the second column of Tables 15. Third, we employ Propensity Score Matching (PSM) to match treated and never-treated hotels.[22] We provide details about this procedure in Online Appendix D. We report the estimates using the matched sample in the last column of Table 15.

City-level time-varying shocks. A related concern is that responding is correlated with city-level time-varying shocks to ratings. To reduce this concern, we estimate Equation 3 by

[22] It is worth noting that, in a generalized DD setup with variation in treatment timing, the treatment group may change every period, making it hard do any meaningful matching. Nevertheless, one can try to attain more balance by matching hotels that eventually respond (treated hotels) to hotels that do not respond (never-treated hotels), as the latter set of hotels is always a part of the control group.

Table 15: Robustness Checks: Accounting for Differences between Responding and Non-responding Hotels

	Hotel-specific Time Trends	Only Responding Hotels	PSM
After Responding	0.079***	0.077***	0.116***
	(0.013)	(0.013)	(0.035)
Avg. Expedia Rating$_{t-1}$	-0.169***	0.264***	0.341***
	(0.022)	(0.036)	(0.071)
Log Number of Expedia Reviews$_{t-1}$	0.028**	-0.011	0.008
	(0.013)	(0.009)	(0.013)
Observations	746464	574252	746464
R^2	0.215	0.169	0.205

Significance levels: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.
Note: The dependent variable is the rating of review i of hotel j at time t. Cluster-robust standard errors (at the individual hotel level) are shown in parentheses. All specifications include hotel and year-month fixed effects.

including the city-year-month fixed effects, effectively allowing for time-varying confounders that vary across cities. We report the results in the first column of Tables 16.

Change in hotel management. Another concern is related to a change in management. Suppose that hotels begin to respond to advertise a recent change in management; then our results may be biased if the new management offers a better service than the old one. To address this concern, we start by identifying hotels that are likely to have experienced a change in management around the time of the first response. We do so by searching for several phrases that indicate managerial changes (e.g., “new management” or “management change”) in a hotel’s responses during its first month since the first response. This strategy is particularly effective because low-rated hotels that underwent a change in management have a strong incentive to indicate this change in their responses to customer reviews. We identify 58 low-rated hotels that likely underwent a change in management around the time of the first response. This suggests that changes in management are not so common for the hotels in our sample. Then, we estimate Equation 3 excluding these 58 hotels. We report the results in the second column of Table 16. We find that the coefficient of interest is very similar to our main estimates reported in the third column of Table 7.

Table 16: Robustness Checks: Controlling for Temporal Shocks, Change in Management, Renovations, and Hotel Price

	City-Month FE	Management Change	Renovations	Price
After Responding	0.099^{***}	0.094^{***}	0.083^{***}	0.097^{***}
	(0.015)	(0.017)	(0.019)	(0.014)
Avg. Expedia Rating$_{t-1}$	0.199^{***}	0.270^{***}	0.269^{***}	0.182^{***}
	(0.026)	(0.034)	(0.037)	(0.029)
Log Number of Expedia Reviews$_{t-1}$	0.005	-0.005	0.003	0.001
	(0.007)	(0.008)	(0.009)	(0.008)
Log Price				-0.175^{***}
				(0.02)
Observations	746325	724146	606817	516058
R^2	0.216	0.203	0.210	0.203

Significance levels: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.
Note: The dependent variable is the rating of review i of hotel j at time t. Cluster-robust standard errors (at the individual hotel level) are shown in parentheses. All specifications include hotel and year-month fixed effects.

Hotel renovations. It is plausible that hotels undergo (or plan to undergo) renovations and begin responding to online reviews as a part of the overall improvement campaign. If this is indeed the case, then the ratings might increase not necessarily because hotels pay attention to the reviews but due to the renovations. To address this concern, we first identify hotels that are likely to have done (or plan to do) renovations around the time of the first response. We do so by searching for text strings that indicate renovations (e.g., “renovate”, “upgrade”) in the responses of a hotel during the first month after beginning to respond. This strategy is particularly effective because low-rated hotels that have done renovations or plan to renovate have a strong incentive to advertise this in their responses to customer reviews. Using this approach, we identify 276 low-rated hotels that indicated renovation around the time of the first response. Then, we estimate Equation 3 by excluding these hotels. We report the results in the third column of Tables 16. We find that the coefficient of interest is similar to our main estimates reported in the third column of Table 7.

Change in prices. Another factor that can affect ratings is price. Suppose hotels reduce their prices or offer discounts and, at the same time, begin to respond to reviews to promote these changes in price. Then, our results could be biased upwards because lower prices could lead to an increase in ratings. To reduce this concern, we obtained price data from STR, a company tracking hotels' economic outcomes. STR tracks the average daily rates (ADR) for 84% of the hotels in our sample.[23] We then estimate Equation 3 including ADR as a control variable. We report these results in the last column of Table 16.

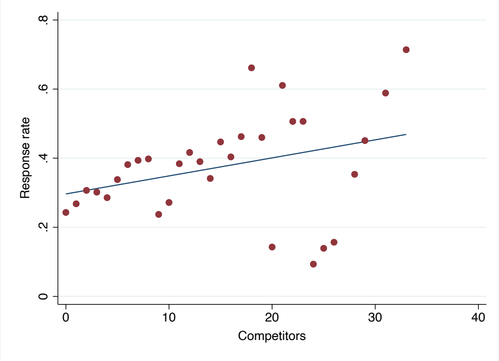


Figure 6: Response rate vs number of competing hotels

Competition. An additional concern is related to competition. Suppose that the likelihood of responding to reviews is driven by how competitive the market is, and more competitive markets lead to higher ratings, then our results could be biased. To address this concern, we start by defining competitors as hotels in the same ZIP code within the same price tier in a given year. We then define the response likelihood of a hotel as the fraction of reviews written after their first management response.

Next, we verify that competition is correlated with hotels' response likelihood. To do so, we plot the distribution of response likelihood by the magnitude of competition. We plot the distribution of response likelihood by the magnitude of competition in Figure 6. The figure indeed suggests that response rate is positively correlated with competition.

We, therefore, proceed to address this concern directly by controlling for the number of competitors, competitors' ratings, and competitors' response rates. We report these results in Table 17.

[23] Some of the hotels for which we have reviews are not tracked by STR.

Table 17: Robustness Checks: Controlling for Competition-related Factors

	(1)	(2)	(3)
After Responding	0.097***	0.098***	0.098***
	(0.016)	(0.018)	(0.017)
Avg. Expedia Rating$_{t-1}$	0.264***	0.296***	0.271***
	(0.030)	(0.040)	(0.033)
Log Number of Expedia Reviews$_{t-1}$	-0.003	-0.003	-0.004
	(0.007)	(0.009)	(0.008)
Competitors	-0.011**		
	(0.006)		
Competitors' Rating		-0.0002	
		(0.016)	
Competitors' Response Rate			-0.012
			(0.021)
Observations	746464	649621	746464
R^2	0.206	0.202	0.206

Significance levels: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.
Note: The dependent variable is the rating of review i of hotel j at time t. Cluster-robust standard errors (at the individual hotel level) are shown in parentheses. All specifications include hotel and year-month fixed effects.

Alternative treatment variables. So far we used a binary treatment variable. In doing so, we implicitly assumed that every hotel responds with roughly the same frequency once it begins responding. However, in practice, we observe considerable variation in the way in which hotels respond to reviews. Intuitively, if responding is correlated with listening as we hypothesized, we should expect that hotels that respond more often are also more likely to keep track of their customer complaints. This, in turn, should have a stronger effect on ratings when compared with hotels that seldom respond.

We test this hypothesis by constructing two measures that track how often a hotel responds to its reviews. The first measure is the *frequency of response*, which we define as the (cumulative) fraction of months in which there is at least one response over the total number of months since the first response. The mean, median, and standard deviation of this measure are 0.5, 0.5, and 0.3 respectively (we present the full distribution of this variable in Figure 7). Note that this measure varies over time because it is updated every month after the first response. The second measure is the *average frequency of response*, which

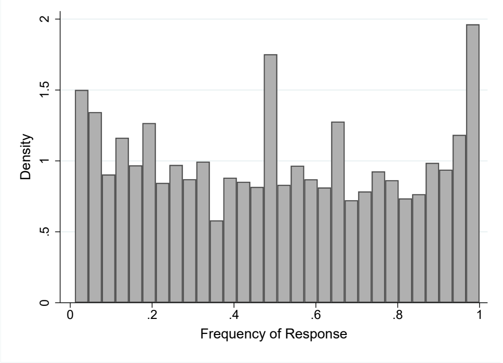


Figure 7: Distribution of Frequency of Response

Table 18: Robustness Checks: Alternative Treatment Variables

	(1)	(2)
Response Frequency	0.138***	0.152***
	(0.020)	(0.024)
Avg. Expedia Rating$_{t-1}$	0.273***	0.273***
	(0.033)	(0.033)
Log Number of Expedia Reviews$_{t-1}$	-0.002	-0.002
	(0.008)	(0.008)
Observations	746464	746464
R^2	0.206	0.206

Significance levels: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.
Note: In Column 1, the independent variable is the frequency of response of hotel j at time t; in Column 2 the independent variable is the average frequency of response of hotel j. Cluster-robust standard errors (at the individual hotel level) are shown in parentheses. All specifications include hotel and year-month fixed effects.

is defined as the average of the frequency of response across post-treatment months. The mean, median, and standard deviation of this measure are 0.48, 0.5, and 0.28, respectively. Note that this measure does not vary over time after the first response.

We report the results of this analysis in Tables 18. Column 1 reports the estimates using the frequency of response; Column 2 shows the estimates using the average frequency of response. Consistent with our hypothesis, we continue to find a positive and statistically significant effect on Expedia ratings, suggesting that the more frequently hotels respond to reviews, the more their ratings increase. Specifically, a low-rated hotel that responds every month sees its rating increase by about 0.14 stars on average.

Sensitivity to Different Sample Periods and Time Windows. As our sample period spans over ten years, it is possible that the positive effect we see is driven by only a few calendar years or is highly sensitive to the treatment window (i.e., length of exposure to the treatment). To check if this is indeed the case, we repeat our main analyses with shorter sample periods and with different before- and after-treatment windows. In column 1 of Table 19, we limit the period to 2009-13; in column 2 we limit the period to 2010-13; in column 3, we limit the period to 2011-13; and in column 4, we limit the period to 2012-13. In addition, we present the results using 12-month and 24-month windows around the time of the first response in Table 20. In all the cases, the estimate is positive and statistically significant, and comparable to that reported in Table 7, column 3 of the paper.

Table 19: Robustness Checks: Different Sample Periods

	2009-13	2010-13	2011-13	2012-13
After Responding	0.087***	0.085***	0.079***	0.077***
	(0.014)	(0.014)	(0.016)	(0.020)
Avg. Expedia Rating$_{t-1}$	0.189***	0.125***	0.092**	-0.072
	(0.039)	(0.041)	(0.046)	(0.053)
Log Number of Expedia Reviews$_{t-1}$	0.016*	0.022**	0.024**	0.021
	(0.009)	(0.009)	(0.011)	(0.014)
Observations	621113	551041	494939	393099
R^2	0.219	0.226	0.232	0.243

Significance levels: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.
Note: The dependent variable is the rating of review i of hotel j at time t. Cluster-robust standard errors (at the individual hotel level) are shown in parentheses. All specifications include hotel and year-month fixed effects.

8 Discussion and Conclusions

In this paper, we demonstrate that online reviews can provide valuable feedback for managers in a dynamic quality environment. Our results show that by listening to reviews and making improvements in quality based on the feedback they receive, hotels can improve their online ratings. We arrive at these results by combining econometrics and machine learning tools. First, we use a difference-in-differences strategy to show that hotels that are more likely

Table 20: Robustness Checks: Different Time Windows around the First Response

	12-month	24-month
After Responding	0.072***	0.089***
	(0.013)	(0.017)
Avg. Expedia Rating$_{t-1}$	0.156***	0.186***
	(0.056)	(0.047)
Log Number of Expedia Reviews$_{t-1}$	0.029***	0.017*
	(0.011)	(0.009)
Observations	345244	477520
R^2	0.232	0.228

Significance levels: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.
Note: The dependent variable is the rating of review i of hotel j at time t. Cluster-robust standard errors (at the individual hotel level) are shown in parentheses. All specifications include hotel and year-month fixed effects.

to pay attention to online reviews increase their ratings more than hotels that do not do so. Second, we reinforce these results by using state-of-the-art ML tools for NLP tasks to analyze the content of the reviews. We show that the average sentiment of issues frequently discussed in the reviews increases after hotels start responding, confirming that the increase in ratings we observe is, at least in part, due to what consumers mention in their feedback. Third, we show that for hotels that are likely to care less about what customers have to say, ratings improve less than hotels that are likely to care more about what their customers have to say.

Our results have important implications for consumers, review platforms, and firms. While prior research has focused on how online reviews help consumers separate high- and low-quality firms, we show that online reviews can help low-quality firms improve their quality, which can increase consumer welfare as the average quality of consumer choices increases. In addition, by realizing that firms take action on their complaints, consumers may be further incentivized to write online reviews. This increase in user engagement on review platforms can lead to a better representation of customer opinions and improve consumer welfare; and review platforms can leverage the increase in user engagement to provide better recommendations.

Prior research on online reviews has provided evidence regarding how ratings on online review platforms can help high-quality firms signal their quality. Our results demonstrate that online review platforms can also create value for low-rated firms by informing them of their shortcomings and giving them an opportunity to redeem themselves in the eyes of the customers. It is also worth noting that while past research has shown how online review platforms help independent and smaller firms generate higher revenue (Hollenbeck 2018), our research shows that, among low-rated hotels, online reviews are leveraged more by bigger and resourceful firms like chain or upscale hotels for quality improvement.

Turning to firms, our results demonstrate that firms use online review platforms as a way to listen to customers, connect with them, and act on the feedback they receive. While the act of obtaining feedback is not new, certain aspects of online review platforms fundamentally change the way in which hotels collect and respond to customer feedback. In the past, firms would collect private feedback from consumers using tools such as surveys; these surveys would generally focus on issues firms thought they had and could improve. However, online reviews provide a source of voluntary feedback that can inform firms about issues they did not know they had. In addition, the public nature of online reviews can magnify these issues, and silence from the management could be perceived as an unwillingness to listen to consumers which, in turn, could lead to a decrease in the firm reputation. Therefore, firms that are not doing so already, should start to pay attention to the reviews if they want to continue to have a good reputation in the eyes of consumers and thus remain competitive.

Despite being around for over two decades and being the focus of a large body of research, online reviews continue to spark interesting questions and debates. In this paper, we identify a simple yet important question with a challenging identification problem. We answer this question by leveraging a difference-in-differences strategy and recent development in NLP tools. We hope that our findings will generate more research in the area of online feedback-based quality improvements.

References

Adamopoulos, Panagiotis, Anindya Ghose, Vilma Todri. 2018. The impact of user personality traits on word of mouth: Text-mining social media platforms. *Information Systems Research* **29**(3) 612–640.

Alam, Ian, Chad Perry. 2002. A customer-oriented new service development process. *Journal of services Marketing* **16**(6) 515–534.

Ang, Desmond. 2021. The effects of police violence on inner-city students. *The Quarterly Journal of Economics* **136**(1) 115–168.

Berger, Jonah, Ashlee Humphreys, Stephan Ludwig, Wendy W Moe, Oded Netzer, David A Schweidel. 2020. Uniting the tribes: Using text for marketing insight. *Journal of Marketing* **84**(1) 1–25.

Berry, Leonard L, A Parasuraman. 1997. Listening to the customer–the concept of a service-quality information system. *MIT Sloan Management Review* **38**(3) 65.

Bertrand, Marianne, Esther Duflo, Sendhil Mullainathan. 2004. How much should we trust differences-in-differences estimates? *The Quarterly journal of economics* **119**(1) 249–275.

Bertschek, Irene, Reinhold Kesler. 2017. Let the user speak: Is feedback on facebook a source of firms' innovation? *ZEW-Centre for European Economic Research Discussion Paper* (17-015).

Burtch, Gordon, Seth Carnahan, Brad N Greenwood. 2018. Can you gig it? an empirical examination of the gig economy and entrepreneurial activity. *Management Science* **64**(12) 5497–5520.

Büschken, Joachim, Greg M Allenby. 2016. Sentence-based text analysis for customer reviews. *Marketing Science* **35**(6) 953–975.

Callaway, Brantly, Pedro HC SantAnna. 2020. Difference-in-differences with multiple time periods. *Journal of Econometrics* .

Cameron, A Colin, Pravin K Trivedi. 2005. *Microeconometrics: methods and applications*. Cambridge university press.

Cer, Daniel, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder. *arXiv preprint arXiv:1803.11175* .

Chakraborty, Ishita, Minkyung Kim, K Sudhir. 2021. Attribute sentiment scoring with online text reviews: Accounting for language structure and missing attributes .

Chevalier, Judith A, Yaniv Dover, Dina Mayzlin. 2018. Channels of impact: User reviews when quality is dynamic and managers respond. *Marketing Science* .

De Chaisemartin, Clément, Xavier d'Haultfoeuille. 2020. Two-way fixed effects estimators with heterogeneous treatment effects. *American Economic Review* **110**(9) 2964–96.

Feng, Juan, Xin Li, Xiaoquan Zhang. 2019. Online product reviews-triggered dynamic pricing: Theory and evidence. *Information Systems Research* **30**(4) 1107–1123.

Gentzkow, Matthew, Jesse M Shapiro, Michael Sinkinson. 2011. The effect of newspaper entry and exit on electoral politics. *American Economic Review* **101**(7) 2980–3018.

Godes, David, José C Silva. 2012. Sequential and temporal dynamics of online opinion. *Marketing Science* **31**(3) 448–473.

Goodman-Bacon, Andrew. 2021. Difference-in-differences with variation in treatment timing. *Journal of Econometrics* .

Griffin, Abbie, John R Hauser. 1993. The voice of the customer. *Marketing science* **12**(1) 1–27.

Gu, Bin, Qiang Ye. 2014. First step in social media: Measuring the influence of online management responses on customer satisfaction. *Production and Operations Management* **23**(4) 570–582.

Hartmann, Jochen, Mark Heitmann, Christian Siebert, Christina Schamp. 2022. More than a feeling: Accuracy and application of sentiment analysis. *International Journal of Research in Marketing* .

Hauser, John R, Don Clausing, et al. 1988. The house of quality .

Herrmann, Andreas, Frank Huber, Christine Braunstein. 2000. Market-driven product and service design: Bridging the gap between customer needs, quality management, and customer satisfaction. *International Journal of production economics* **66**(1) 77–96.

Hollenbeck, Brett. 2017. The economic advantages of chain organization. *The RAND Journal of Economics* **48**(4) 1103–1135.

Hollenbeck, Brett. 2018. Online reputation mechanisms and the decreasing value of chain affiliation. *Journal of Marketing Research* **55**(5) 636–654.

Kaulio, Matti A. 1998. Customer, consumer and user involvement in product development: A framework and a review of selected methods. *Total quality management* **9**(1) 141–149.

Krishnan, Viswanathan, Karl T Ulrich. 2001. Product development decisions: A review of the literature. *Management science* **47**(1) 1–21.

Lee, Thomas Y, Eric T Bradlow. 2011. Automated marketing research using online customer reviews. *Journal of Marketing Research* **48**(5) 881–894.

Liu, Xiao, Dokyun Lee, Kannan Srinivasan. 2018. Large scale cross-category analysis of consumer review content on sales conversion leveraging deep learning .

Luca, Michael. 2016. Reviews, reputation, and revenue: The case of yelp. com. *Com (March 15, 2016). Harvard Business School NOM Unit Working Paper* (12-016).

Matzler, Kurt, Hans H Hinterhuber. 1998. How to make product development projects more successful by integrating kano's model of customer satisfaction into quality function deployment. *Technovation* **18**(1) 25–38.

Mayzlin, Dina, Yaniv Dover, Judith Chevalier. 2014. Promotional reviews: An empirical investigation of online review manipulation. *American Economic Review* **104**(8) 2421–55.

Mikolov, Tomas, Kai Chen, Greg Corrado, Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* .

Mousavi, Reza, Bin Gu. 2019. The impact of twitter adoption on lawmakers voting orientations. *Information Systems Research* **30**(1) 133–153.

Netzer, Oded, Ronen Feldman, Jacob Goldenberg, Moshe Fresko. 2012. Mine your own business: Market-structure surveillance through text mining. *Marketing Science* **31**(3) 521–543.

Prager, Elena, Matt Schmitt. 2021. Employer consolidation and wages: Evidence from hospitals. *American Economic Review* **111**(2) 397–427.

Proserpio, Davide, Isamar Troncoso, Francesca Valsesia. 2021. Does gender matter? the effect of management responses on reviewing behavior. *Marketing Science* .

Proserpio, Davide, Georgios Zervas. 2017. Online reputation management: Estimating the impact of management responses on consumer reviews. *Marketing Science* **36**(5) 645–665.

Puranam, Dinesh, Vrinda Kadiyali, Vishal Narayan. 2021. The impact of increase in minimum wages on consumer perceptions of service: A transformer model of online restaurant reviews. *Marketing Science* **40**(5) 985–1004.

Schaffhausen, Cory R, Timothy M Kowalewski. 2015. Large-scale needfinding: methods of increasing user-generated needs from large populations. *Journal of Mechanical Design* **137**(7) 071403.

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

Sullivan, Lawrence P. 1986. Quality function deployment. *Quality Progress (ASQC)* 39–50.

Sun, Liyang, Sarah Abraham. 2020. Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics* .

Timoshenko, Artem, John R Hauser. 2019. Identifying customer needs from user-generated content. *Marketing Science* **38**(1) 1–20.

Tirunillai, Seshadri, Gerard J Tellis. 2012. Does chatter really matter? dynamics of user-generated content and stock performance. *Marketing Science* **31**(2) 198–215.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*. 5998–6008.

Wang, Yang, Alexander Chaudhry. 2018. When and how managers responses to online reviews affect subsequent reviews. *Journal of Marketing Research* **55**(2) 163–177.

Wang, Yang, Alexander Chaudhry, Amit Pazgal. 2020. Do online reviews improve product quality? evidence from hotel reviews on travel sites. *Available at SSRN 3238510* .

Zhang, Yi, Camilla Vásquez. 2014. Hotels' responses to online reviews: Managing consumer dissatisfaction. *Discourse, Context & Media* **6** 54–64.

Online Appendix

A Heterogeneous Effects: Star-rating Analyses

In this section, we consider heterogeneous quality improvements across different types of hotels. Specifically, we study how the effect varies with (i) hotel operation type, i.e., chain and independent hotels, and (ii) price tier (categorized as low-, mid-, and high-tier based on the hotel class information provided by Tripadvisor).

A.1 Hotel Operation

We hypothesize that the effect on ratings should be higher for chain hotels because of their supply-side (cost) and demand-side (reputation) advantages. Chain hotels are more likely to incorporate customer feedback as they incur lower operational marginal costs due to economies of scale. Chain hotels also generate a greater premium due to a higher reputation on online platforms (Hollenbeck 2017, 2018). Moreover, the reputational costs of disregarding customer feedback on online review platforms could be higher for chain hotels, spurring them to act on customer feedback.

To test this hypothesis, we estimate the following model specification:

$$Stars_{ijt}^{E} = \beta_1 \text{After Responding}_{ijt}^{TA} + \beta_2 \text{After Responding}_{ijt}^{TA} \times \text{Chain}_j + \text{X}_{jt}^{\prime}\gamma + \alpha_j + \tau_t + \epsilon_{ijt}, \quad (4)$$

where everything is as in Equation 3, and Chain_j is an indicator of whether hotel j is part of a chain. β_2 is the coefficient of interest, which measures the differential impact of listening to reviews for chain hotels. We report the estimates of Equation 4 in the first column of Table 21. In line with our hypothesis, we observe that the coefficient of interest is positive and statistically significant, suggesting that chain hotels do increase their quality more than independent hotels. In Table 22, we present the results for hotels that were previously high-rated. The results suggest that both independent and chain hotels fail to improve their ratings.

A.2 Hotel Price Tier

Here we analyze whether hotels with more resources (such as high-end hotels) and therefore with a better ability to identify customer needs see a larger gain in average rating than those hotels that are relatively more constrained in terms of resources (such as budget hotels). For example, it is far more likely that high-end hotels hire reputation management specialists to swiftly identify issues from customer feedback and respond to those issues on review platforms than low-end hotels. Further, the ability to close the feedback loop by leveraging resources and access may be higher for high-end hotels. To verify our hypothesis, we estimate a model similar to Equation 4, but now interact the main variable of interest with a discrete variable that indicates whether a hotel is low-tier, mid-tier, or high-tier. We report the estimates in the second column of Table 21. As expected, we find that, compared to low- and mid-tier hotels, high-tier hotels experience a bigger improvement in their ratings.

In Table 22, we present the results for hotels that were previously high-rated. The results suggest that low-tier high-rated hotels experience a decline in ratings while the higher-tier hotels show a slight improvement in ratings.

Table 21: Heterogeneous Effects for Low-rated Hotels: Hotel Operation and Price Tier

	(1)	(2)
After Responding	0.061**	0.080***
	(0.025)	(0.020)
After Responding × Chain	0.065***	
	(0.025)	
After Responding × Mid-tier		0.004
		(0.025)
After Responding × High-tier		0.084***
		(0.026)
Avg. Expedia Rating$_{t-1}$	0.268***	0.273***
	(0.033)	(0.033)
Log Number of Expedia Reviews$_{t-1}$	-0.003	-0.002
	(0.008)	(0.007)
Observations	746464	746464
R^2	0.206	0.206

Significance levels: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.
Note: The dependent variable is the rating of review i of hotel j at time t. Cluster-robust standard errors (at the individual hotel level) are shown in parentheses. All specifications include hotel and year-month fixed effects.

Table 22: Heterogeneous Effects for High-rated Hotels: Hotel Operation and Price Tier

	(1)	(2)
After Responding	0.017	-0.048***
	(0.022)	(0.012)
After Responding × Chain	-0.012	
	(0.024)	
After Responding × Mid-tier		0.073***
		(0.020)
After Responding × High-tier		0.078***
		(0.021)
Avg. Expedia Rating$_{t-1}$	0.265***	0.260***
	(0.024)	(0.024)
Log Number of Expedia Reviews$_{t-1}$	0.001	0.001
	(0.006)	(0.006)
Observations	1104851	1104851
R^2	0.147	0.147

Significance levels: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.
Note: The dependent variable is the rating of review i of hotel j at time t. Cluster-robust standard errors (at the individual hotel level) are shown in parentheses. All specifications include hotel and year-month fixed effects.

B Robustness Checks: Entity Sentiment Analyses

In Table 23, we replicate Table 10 but after excluding reviews that do not mention the entities.

In Table 24, we replicate Table 11 but account for the average pre-treatment sentiment of the entity.

Table 23: Robustness Checks: Excluding Reviews that do not Mention the Entity

	1st Most	2nd Most	3rd Most	4th Most	5th Most
After Responding	0.1109***	0.1301***	0.1035***	0.1163***	0.1008***
	(0.0197)	(0.0205)	(0.0162)	(0.0180)	(0.0159)
Avg. Expedia Rating$_{t-1}$	0.0661**	0.0015	0.1068***	0.0325	0.0400
	(0.0311)	(0.0312)	(0.0312)	(0.0326)	(0.0309)
Log Number of Expedia Reviews$_{t-1}$	0.0430***	-0.0090	0.0014	0.0180	0.0004
	(0.0113)	(0.0127)	(0.0120)	(0.0115)	(0.0104)
Observations	10227	9376	11869	11498	12238
R^2	0.2455	0.1419	0.1666	0.1492	0.1655

Significance levels: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.
Note: The dependent variable is the sentiment score of a top-5 entity in review i of hotel j at time t. Cluster-robust standard errors (at the individual hotel level) are shown in parentheses. All specifications include hotel and year-month fixed effects.

Table 24: Robustness Checks: Heterogeneous Effects Accounting for the average pre-treatment Sentiment of the Entity

	1st Most	2nd Most	3rd Most	4th Most	5th Most
After Responding	-0.0158***	-0.0052	-0.0011	-0.0036**	-0.0007
	(0.0028)	(0.0039)	(0.0019)	(0.0017)	(0.0016)
After Responding × Pre NegMentions	0.3438***	0.2321**	0.1153**	0.2306***	0.1561***
	(0.0417)	(0.0954)	(0.0550)	(0.0413)	(0.0568)
After Responding × Pre Sentiment	-0.0972***	-0.0589***	-0.0459***	-0.0388***	-0.0286***
	(0.0110)	(0.0166)	(0.0091)	(0.0064)	(0.0068)
Avg. Expedia Rating$_{t-1}$	0.0049	0.0010	0.0090***	0.0033*	0.0038*
	(0.0041)	(0.0032)	(0.0027)	(0.0020)	(0.0021)
Log Number of Expedia Reviews$_{t-1}$	0.0019	-0.0016	-0.0003	0.0003	0.0002
	(0.0016)	(0.0014)	(0.0009)	(0.0008)	(0.0006)
Observations	91938	104244	167853	179953	217555
R^2	0.0523	0.0171	0.0172	0.0120	0.0142

Significance levels: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.
Note: The dependent variable is sentiment score of a top-5 entity in review i of hotel j at time t. Cluster-robust standard errors (at the individual hotel level) are shown in parentheses. All specifications include hotel and year-month fixed effects.

C Additional Robustness checks: Entity Sentiment Analyses

In this section, we present the same set of robustness checks discussed in Section 7, but for the analyses where the outcome is the sentiment of the most frequently mentioned entity.

In Table 25, we show that the results are robust to the inclusion of hotel-specific time trends (column 1), removing hotels that did not respond to reviews (column 2), and considering only hotels matched using Propensity Score Matching (column 3).

In Table 26, we show that the results are robust to the inclusion of city-month FE (column 1), removing hotels that experience a change in management (column 2), excluding hotels that likely went through renovations (column 3), and controlling for hotel average daily rates (column 4).

In Table 27, we show that the results are robust to the inclusion of competition-related factors.

In Table 28, we show that the results are robust to the use of alternative treatment variables.

In Table 29, we show that our results are robust to considering shorter sample periods.

In Table 30, we show that our results are robust to considering different time windows around the time of the first response of each hotel.

Table 25: Robustness Checks: Accounting for Differences between Responding and Non-responding Hotels

	Hotel-specific	Only Responding	PSM
	Time Trends	Hotels	
After Responding	0.0173***	0.0165***	0.0160***
	(0.0028)	(0.0028)	(0.0028)
Avg. Expedia Rating$_{t-1}$	-0.0015	0.0046	0.0081
	(0.0048)	(0.0046)	(0.0057)
Log Number of Expedia Reviews$_{t-1}$	0.0042	0.0058***	0.0043**
	(0.0027)	(0.0021)	(0.0019)
Observations	91938	76995	90706
R^2	0.0580	0.0568	0.0460

Significance levels: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.
Note: The dependent variable is the sentiment score of the most frequently mentioned problematic entity in review i of hotel j at time t. Cluster-robust standard errors (at the individual hotel level) are shown in parentheses. All specifications include hotel and year-month fixed effects.

Table 26: Robustness Checks: Controlling for Temporal Shocks, Change in Management, Renovations, and Hotel Price

	City-Month	Management	Renovations	Price
	FE	Change		
After Responding	0.0157***	0.0158***	0.0173***	0.0163***
	(0.0026)	(0.0026)	(0.0030)	(0.0027)
Avg. Expedia Rating$_{t-1}$	0.0065	0.0054	0.0088**	0.0015
	(0.0048)	(0.0040)	(0.0044)	(0.0046)
Log Number of Expedia Reviews$_{t-1}$	0.0056***	0.0044**	0.0045**	0.0065***
	(0.0019)	(0.0018)	(0.0019)	(0.0020)
Log Price				-0.0025
				(0.0047)
Observations	91513	89022	80084	73387
R^2	0.0964	0.0506	0.0373	0.0524

Significance levels: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.
Note: The dependent variable is the sentiment score of the most frequently mentioned problematic entity in review i of hotel j at time t. Cluster-robust standard errors (at the individual hotel level) are shown in parentheses. All specifications include hotel and year-month fixed effects.

Table 27: Robustness Checks: Controlling for Competition-related Factors

	(1)	(2)	(3)
After Responding	0.0169***	0.0155***	0.0169***
	(0.0026)	(0.0029)	(0.0026)
Avg. Expedia Rating$_{t-1}$	0.0073*	0.0087*	0.0077*
	(0.0042)	(0.0048)	(0.0042)
Log Number of Expedia Reviews$_{t-1}$	0.0048***	0.0056***	0.0047***
	(0.0018)	(0.0020)	(0.0018)
Competitors	-0.0010		
	(0.0006)		
Competitors' Rating		0.0032	
		(0.0033)	
Competitors' Response Rate			0.0013
			(0.0037)
Observations	91938	77725	91938
R^2	0.0501	0.0557	0.0501

Significance levels: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.
Note: The dependent variable is the sentiment score of the most frequently mentioned problematic entity in review i of hotel j at time t. Cluster-robust standard errors (at the individual hotel level) are shown in parentheses. All specifications include hotel and year-month fixed effects.

Table 28: Robustness Checks: Alternative Treatment Variables

	(1)	(2)
Response Frequency	0.0114***	0.0156***
	(0.0039)	(0.0052)
Avg. Expedia Rating$_{t-1}$ 0.0074*	0.0073*	
	(0.0043)	(0.0042)
Log Number of Expedia Reviews$_{t-1}$	0.0058***	0.0055***
	(0.0018)	(0.0018)
Observations	91938	91938
R^2	0.0495	0.0496

Significance levels: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.
Note: In Column 1, the independent variable is the frequency of response of hotel j at time t; in Column 2 the independent variable is the average frequency of response of hotel j. Cluster-robust standard errors (at the individual hotel level) are shown in parentheses. All specifications include hotel and year-month fixed effects.

Table 29: Robustness Checks: Different Sample Periods

	2009-13	2010-13	2011-13	2012-13
After Responding	0.0176***	0.0197***	0.0188***	0.0153***
	(0.0030)	(0.0036)	(0.0040)	(0.0050)
Avg. Expedia Rating$_{t-1}$	0.0050	0.0061	0.0184	-0.0037
	(0.0066)	(0.0086)	(0.0117)	(0.0141)
Log Number of Expedia Reviews$_{t-1}$	0.0063***	0.0078**	0.0049	0.0054
	(0.0022)	(0.0031)	(0.0034)	(0.0042)
Observations	78535	69435	62402	49614
R^2	0.0558	0.0602	0.0629	0.0719

Significance levels: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.
Note: The dependent variable is the sentiment score of the most frequently mentioned problematic entity in review i of hotel j at time t. Cluster-robust standard errors (at the individual hotel level) are shown in parentheses. All specifications include hotel and year-month fixed effects.

Table 30: Robustness Checks: Different Time Windows around the First Response

	12-month	24-month
After Responding	0.0180***	0.0169***
	(0.0030)	(0.0029)
Avg. Expedia Rating$_{t-1}$	0.0061	0.0093
	(0.0081)	(0.0069)
Log Number of Expedia Reviews$_{t-1}$	0.0069**	0.0079***
	(0.0030)	(0.0023)
Observations	39034	57526
R^2	0.058	0.060

Significance levels: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.
Note: The dependent variable is the sentiment score of the top entity in review i of hotel j at time t. Cluster-robust standard errors (at the individual hotel level) are shown in parentheses. All specifications include hotel and year-month fixed effects.

D Propensity Score Matching details

We matched treated and never-treated hotels using the following variables: whether the hotel is part of a chain, price tier (low, mid, or high), city, and whether we classify the hotel as high-rated. We report the balance check in Table 31.

Table 31: Difference in Covariates Between Responding and Non-responding Hotels after PSM

Variable	Treated	Never-treated	P-value
Chain	0.77	0.76	0.08
Price tier	0.70	0.69	0.30
City (Index 1-50)	25.7	25.8	0.64
High-rated	0.60	0.60	1.00