Online reputation management: Estimating the impact of management responses on consumer reviews†

Davide Proserpio
Computer Science Department
Boston University

Georgios Zervas
School of Management
Boston University

January 31, 2015

Abstract

Failure to meet a consumer’s expectations can result in a negative review, which can have a lasting, damaging impact on a firm’s reputation, and its ability to attract new customers. To mitigate the reputational harm of negative reviews many firms now publicly respond to them. How effective is this reputation management strategy in improving a firm’s reputation? We empirically answer this question by exploiting a difference in managerial practice across two hotel review platforms, TripAdvisor and Expedia: while hotels regularly respond to their TripAdvisor reviews, they almost never do so on Expedia. Based on this observation, we use difference-in-differences to identify the causal impact of management responses on consumer ratings by comparing changes in the TripAdvisor ratings of a hotel following its decision to begin responding against a baseline of changes in the same hotel’s Expedia ratings. We find that responding hotels, which account for 56% of hotels in our data, see an average increase of 0.12 stars in the TripAdvisor ratings they receive after they start responding. Moreover, we show that this increase in ratings does not arise from hotel quality investments. Instead, we find that the increase is consistent with a shift in reviewer selection: consumers with a poor experience become less likely to leave a negative review when hotels begin responding.

†We thank Frederic Brunel, John Byers, Sharon Goldberg, Tim Simcoe, and Greg Stoddard for helpful comments and discussion.
1 Introduction

User-generated online reviews have been continuously gaining credibility in the eyes of consumers, and today they are an essential component of the consumer decision making process (Chevalier and Mayzlin 2006, Luca 2011). With the popularity and reach of online review platforms growing rapidly, firms are under increasing pressure to maintain a flawless online reputation. While investing in improved products and services can result in better ratings, inevitably firms do experience failures that lead to negative reviews. Dealing with unfavorable reviews is challenging because, unlike offline word-of-mouth, they persist online, and can cause indelible damage to a firm’s reputation. Given that firms can neither selectively delete reviews, nor opt-out from being reviewed altogether, it is essential for managers to develop reputation management strategies that can dampen the reputational shock of negative reviews.

Current reputation management practice encompasses various strategies that vary in their efficacy, and adherence to legal and ethical norms. These strategies range from outright review fraud (Mayzlin et al. 2013, Luca and Zervas 2013), to incentivizing consumers to leave reviews in exchange for perks, to taking legal action against consumers who leave negative reviews, and to using non-disparagement clauses in sales contracts that stipulate fines if consumers write negative reviews. Meanwhile, technological advancement in detecting fraudulent or incentivized reviews, enforcement of false advertising regulations against those who commit review fraud, and emerging legislation aiming to protect consumer free speech online have created an environment where these activities carry significant legal, and financial risk for dubious reward.

In this climate, the practice of publicly responding to consumer reviews has emerged as an alternative reputation management strategy that is legal, endorsed by review platforms, and widely adopted by managers. A management response takes the form of an open-
ended piece of text that is permanently displayed beneath the review it addresses. Unlike the review itself, the response does not carry a rating, and it doesn’t contribute in the calculation of a firm’s average rating. While review platforms ensure that responses meet certain basic standards (such as avoiding offensive language) they allow any firm to respond to any reviewer. Most major review platforms, including TripAdvisor and Yelp, allow firms to respond to their reviews. Yet, despite management responses now being commonplace, their efficacy in recovering a firm’s reputation remains an open question. Our research aims to fill this gap.

In this paper, we focus on the Texas hospitality industry, and we estimate the causal impact of management responses on hotel reputation. We show that, on average, responding hotels see a consistent increase of 0.12 stars in their consumer ratings after they start using management responses. Because TripAdvisor and other review platforms round average ratings to the nearest half-star changes even smaller than 0.12 stars can have material impact. For example, if a 4.24-star hotel can cross the 4.25-star threshold it will see its rating jump by half a star, which can in turn cause a substantial increase in revenue. For restaurants, Luca (2011) finds that a one-star increase corresponds to 5%-9% increased revenue. In our data, 27% of responding hotels were able to increase their rounded rating by at least half a star within 6 months of their first management response. By comparison, only 16% of the same hotels were able to increase their rounded rating in the period starting one year before their first management response and ending 6 months later.

To ascribe a causal interpretation to our findings, we need to address two main identification challenges. The first is hotel self-selection into using management responses. In our data, hotels with higher ratings tend to respond to consumer reviews more often than hotels with lower ratings. Therefore, between hotel comparisons can be biased. The second, and more challenging to address, concern is that a hotel’s decision to begin responding to consumer reviews can be correlated with unobserved changes, such as service improvements and renovations, the hotel made to avoid further negative reviews. According to a recent
New York Times article, hotels commonly use online reviews as a guide for renovations.\(^2\) Therefore, increased ratings following a management response can simply reflect an effort by hotel management to fix the problem that was causing the negative reviews in the first place rather than any direct impact of the management responses themselves.

We address these identification concerns by exploiting a difference in managerial practice between two major review platforms, TripAdvisor and Expedia: while hotels frequently respond to TripAdvisor reviews, the same hotels rarely do so on Expedia. To explain this difference, we point out that unlike TripAdvisor, which is in the business of collecting and disseminating consumer reviews, Expedia is primarily a web travel agency (Mayzlin et al. (2013) make the same point.) Comparing the information available for any given hotel on the two sites highlights this distinction: while TripAdvisor prominently displays a hotel’s reviews, Expedia displays a booking form, prices for various room types, and the hotel’s average rating – individual reviews are only available on a secondary page accessible by following a link on the hotel’s main Expedia page. As a result, in our data, 31.5% of TripAdvisor reviews carry a response, compared to 2.2% on Expedia.

This difference between the two platforms suggests that a TripAdvisor user is more likely to notice that a hotel tends to respond to its reviews that an Expedia user. In turn, knowledge that a hotel monitors and responds to its reviews can affect TripAdvisor users’ reviewing behavior both in terms of the decision to leave a review, and the review’s content. We build upon this observation to estimate the causal impact of management responses using a difference-in-differences (DD) identification strategy. Focusing on hotels that respond on TripAdvisor, but do not respond on Expedia – the majority of properties in our data – we compare changes in the TripAdvisor ratings of any given hotel following its decision to begin responding against a baseline of changes in the same hotel’s Expedia ratings over the same period of time. This strategy addresses both identification challenges we highlighted above. First, by only relying on within-hotel variation for identification, we eliminate the concern of

bias arising from unobserved, time-invariant heterogeneity between hotels. Second, we can account for the confounding effects of unobserved hotel improvements using Expedia ratings as a baseline. Expedia ratings are a credible control group because it is unlikely that unobserved hotel improvements (or, other changes for that matter) will only impact TripAdvisor reviewers, leaving Expedia reviewers unaffected. Concretely, our identification assumption is that any responding hotel’s TripAdvisor and Expedia ratings would have followed parallel trends had the hotel not started responding. While in general this assumption in untestable, the panel nature of our data allows us to partially verify its validity: we show that for a long pre-treatment period – *i.e.*, the period preceding each hotel’s first management response – TripAdvisor and Expedia ratings moved in parallel, lending plausibility to the assumption that this parallel trend would have extended to the post-responses period in the absence of management responses. Overall, our analysis suggests that management responses positively impact hotel reputation. Moreover, this improvement in reputation is a direct outcome of the responses themselves, and not of unobserved improvements or renovations that could also increase consumer satisfaction.

We next turn our attention to establishing the robustness of this finding. In particular, one limitation of using Expedia ratings as a control group relates to unobserved heterogeneity between the two review platforms. For example, unobserved changes in review collection methodology, such as a simple change to a review submission form, could lead to endogeneity by affecting consumer ratings on one platform, but not the other. To avoid bias due to unobserved review platform differences, we repeat our DD analysis using the ratings of non-responding TripAdvisor hotels as a baseline. Because in this analysis the treatment and control groups are composed of different hotel units, we combine DD with matching to further reduce endogeneity concerns (Heckman and Navarro-Lozano 2004). We use the coarsened exact matching (CEM) approach of Iacus et al. (2011) to construct an estimation sample that matches treated and control hotels based on their city, price category, and organization (chain, franchise or independent). We find the DD estimate based on the CEM-matched
sample to be similar to our previous analysis, suggesting that our results are not driven by unobserved review platform biases. As a final robustness check, we combine the two DD strategies we outlined above to identify the causal effect of management responses using a difference-in-difference-in-differences (DDD) design. This strategy employs the full sample of TripAdvisor and Expedia ratings, including both responding and non-responding hotels, and requires even weaker assumptions: no transient shock that differentially affects the cross-platform differences in the ratings of responding hotels, compared to a baseline of cross-platform differences in the ratings for non-responding hotels. Our DDD estimate is 0.11 stars, which is almost identical to our DD estimate. The robustness of these additional analyses increases our confidence in having captured the causal impact of management responses on hotel reputation.

Having estimated the extent to which management responses improve hotel reputation, we next turn our attention to understanding the mechanism underlying this effect. We explore two hypotheses to explain our findings. First, drawing from the service failure and recovery literature (e.g., Tax et al. (1998), Smith et al. (1999), McCollough et al. (2000)) we hypothesize that management responses encourage consumers who left negative reviews to return, give hotels a second try, and possibly leave a fresh, positive review. We find some limited evidence that management responses aid in service recovery: in our data, negative reviewers who received a management response were more likely to return to TripAdvisor and leave a positive review than reviewers who did not receive a response to their initial review. However, the number of such reviews by returning consumers is too small (1.3% of all TripAdvisor reviews) to adequately explain the increase in ratings of responding hotels.

The second theory we investigate to explain our results is that management responses shift reviewer selection towards reviewers that are less stringent. This hypothesis is motivated by a number of studies (e.g., Dellarocas and Wood (2008), Bolton et al. (2013)), which have shown that in settings where both parties participating in a trade (in our case hotels and their guests) can review each other, negative ratings are underreported. One explanation for
the underreporting of negative reviews is fear of reciprocation: agents with a poor experience choose not leave a negative review for the other party to avoid inviting a negative review back. We find evidence in our data consistent with this hypothesis: following a hotel’s decision to begin responding, the reviewers it attracts are inherently more positive in their evaluations. We arrive at this result in two steps. First, we estimate reviewer inherent positivity using a model that predicts each reviewer’s TripAdvisor rating for the average TripAdvisor business. Reviewers with higher predicted ratings for the average business are less stringent. Next, we show that reviewers in the post-responses period would rate the average firm 0.09 stars higher than reviewers in the pre-responses period, suggesting a shift in reviewer population towards less stringent reviewers. The argument that management responses are a signaling mechanism meant to shape the behavior of future TripAdvisor reviewers is also consistent with the response patterns we observe: hotels in our data respond with the same frequency to positive and negative reviews, and they leave responses to reviews that are many years old. Responses to positive or old reviews are unlikely to be part of a service recovery effort, and managers likely leave them to signal their intention of monitoring and responding to consumer reviews.

We organize the remainder of the paper as follows. In Section 2 we outline the empirical framework we use to identify the causal impact of management responses on hotel ratings. Next, in Section 3 we describe the dataset we use, and provide descriptive statistics. In Section 4 we present our results, and in Section 5 we provide theoretical support for our findings. Finally, in Section 6 we discuss the managerial implications of our results, some limitations of our study, and potential directions for future research.

2 Empirical strategy

Our goal is in this paper is to estimate the causal impact of management responses on the ratings of hotels that respond to reviews. This quantity is an average treatment effect on the
treated (ATT), and it is only defined for hotels that have elected to respond to TripAdvisor reviewers. Therefore, it is not necessarily equal to the average treatment effect (ATE), which is the effect management responses would have had on the TripAdvisor ratings of a randomly chosen hotel. To motivate our empirical strategy, we consider an exogenous intervention that would allow us to estimate the ATT. With access to the TripAdvisor platform, we would randomly assign TripAdvisor visitors into one of two conditions: a treatment group exposed to a version of the site that displays management responses (i.e., the current TripAdvisor site), and a control group exposed to a version modified to omit management responses, but identical otherwise. Then, using counterfactual notation, for any responding hotel \( i \) the ATT is

\[
E(Y_{i1} - Y_{i0}|D = 1)
\]

where \( Y_{i1} \) is a TripAdvisor rating for hotel \( i \) from the treatment condition, \( Y_{i0} \) is a TripAdvisor 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.

The key challenge presented by our lack of experimental data is that we do not observe the counterfactual ratings \( Y_{i0} \) that consumers would have submitted had they not been exposed to management responses; this a quantity that we need to identify \( E(Y_{i0}|D = 1) \). To address this identification challenge we need to construct an appropriate control group out of our non-experimental data to stand in for \( Y_{i0} \).

### 2.1 Identification strategy

A first solution, which exploits the panel nature of our data, is to use the ratings of hotel \( i \) submitted prior to its first management response as a control group. Using the superscripts \( pre \) and \( post \) for ratings submitted before and after hotel \( i \) began responding, the required assumption to identify the ATT is \( E(Y_{i0}^{pre}|D = 1) = E(Y_{i0}^{post}|D = 1) \).\(^3\) This assumption

\(^3\)For ease of presentation, we describe our identification strategy in terms of two periods, before and after treatment, but its extension to a setting with multiple pre and post periods is straightforward.
is unlikely to hold, leading to endogeneity in our estimation. The key threat to validity is that hotels often use management responses to advertise improvements they have made following a poor review, and therefore, increased ratings following a management response can be the result of these improvements, rather than the outcome of consumer exposure to the management response itself.

A second solution to the identification challenge is based on the observation that most hotels that respond to their TripAdvisor reviews do not respond to their reviews on Expedia. Therefore, in principle, we could use the Expedia ratings of hotel \( i \) in place of the unobserved counterfactual ratings \( Y_{i0} \). Denoting Expedia ratings by \( Z \), the necessary identification condition is \( E(Y_{i0}|D = 1) = E(Z_{i0}|D = 1) \), and it is also unlikely to hold. The endogeneity issue arising in this case is that TripAdvisor and Expedia reviewers are likely to differ in unobservable ways that determine their ratings. For example, in Table 2, we show that the average hotel rating on TripAdvisor is 0.3 stars lower than on Expedia, i.e., Expedia reviewers report greater levels of satisfaction.

In this paper, we combine the above two approaches in a difference-in-differences (DD) identification strategy, which requires weaker assumptions. We proceed in two steps: first, we construct a matched-control for each hotel’s TripAdvisor ratings using the same hotel’s ratings on Expedia; then, we compare post-treatment differences in the hotel’s TripAdvisor ratings against a baseline of post-treatment differences in same hotel’s Expedia ratings. Formally stated, our main identification assumption is

\[
E(Y_{i0}^{post} - Y_{i0}^{pre} | D = 1, X) = E(Z_{i0}^{post} - Z_{i0}^{pre} | D = 0, X). \tag{1}
\]

This is the so-called parallel-trends assumption of DD models, and it is weaker than both assumptions stated above. It states that, conditional on observed characteristics \( X \), differences in (potential) outcomes do not depend on whether a unit was treated, or not. Crucially, it allows both for platform-independent transient shocks to hotel ratings arising
from quality investments that coincide with management responses, as well as time-invariant cross-platform differences in hotel ratings. Another important advantage of this identification strategy is that we can partially test its underlying assumptions by comparing the pre-treatment ratings trends of treated and control units. We return to this point in Section 4.1, where we show that pre-treatment trends are indeed parallel, thereby providing evidence in support of our main identification assumption. This is our preferred identification strategy, and we will refer to it as cross-platform DD to highlight its use of hotel ratings from both TripAdvisor and Expedia.

Matching and DD have been successfully combined to identify causal effects in a number of studies (e.g., Singh and Agrawal (2011), Azoulay et al. (2013), and Kovács and Sharkey (2014).) The matching step is taken to ensure the compatibility of treated and control units in terms of their potential outcomes, which is the main DD identification assumption. An important distinction of our matching sample with respect to papers that use a similar empirical strategy is that the treated and control units in our study are not just similar with respect to some observable properties, but they are in fact exactly same hotel unit. This exact one-to-one matching procedure almost entirely eliminates the key concern of the comparability among treated and control hotels. This is important because as recognized in Imbens (2004) stringent matching criteria reduce concerns about endogeneity-related biases.

### 2.2 Threats to cross-platform identification, and alternatives

Our cross-platform DD identification strategy is robust to review-platform independent, transitory shocks to hotel ratings. However, unobserved platform-specific shocks to hotel ratings whose timing is correlated with management responses can bias our estimation. Here we describe an identification strategy to mitigate this concern.

Our alternative identification strategy exploits the fact that about half of all hotels do not respond to their TripAdvisor reviews to construct a control group of hotel ratings from within the same review platform. In place of the 1-1 matched Expedia control group, we
now use the TripAdvisor ratings of non-responding hotels. An issue arising from the absence of a natural 1-1 match is that there is no simple way to define a post-treatment period for untreated units. Therefore, instead of differencing with respect to the treatment clock as we did in Equation 1, we now difference with respect to calendar time. We refer to this alternative identification strategy as within-platform DD. Denoting successive calendar months by the superscripts $t$ and $t + 1$, our main identification assumption is

$$E(Y_{i0}^{t+1} - Y_{i0}^{t} | D = 1, X) = E(Y_{i0}^{t+1} - Y_{i0}^{t} | D = 0, X).$$

The key advantage of this identification strategy is that it eliminates the possibility of bias due to unobserved cross-platform differences. However, bias could now arise from unobserved, time-varying confounders that are correlated with the timing of the treatment. To guard against this concern we use matching methods to ensure that treated and control units are similar in their observable dimensions. The assumption underlying matching procedures is that matching treated and control units along observed characteristics can significantly reduce (but not completely rule out) endogeneity concerns. (Heckman and Navarro-Lozano 2004)

In this paper, we use the Coarsened Exact Matching (CEM) procedure of Iacus et al. (2011). CEM begins with a choice of variables to be used for matching treated and control units. Continuous variables, or variables that take on too many values, need to be coarsened, that is to say, discretized to a smaller set of possible values. The data is then stratified based on these variables, and strata that do not contain both treated and control units are discarded. The procedure entails a trade-off between the number of variables that are used for stratification, and the amount of data that has to be discarded because no match can be found, a problem often referred to as the curse of dimensionality. Once matching is performed, the observations are reweighed to correct any imbalance in the number of treated and control units within each stratum. We prefer CEM over other matching approaches
because, in addition to its statistical advantages (Iacus et al. 2011), transparency, and ease of implementation, it is a natural fit for our dataset, which is rich in categorical variables that we can use for stratification. CEM does not prescribe a particular selection of stratification criteria; this choice is primarily guided by domain knowledge. In our setting, we choose to match hotels based on their city, price (budget, economy, midprice, upscale, or luxury), and operation (independent, franchise, or chain). We refer to this strategy as CEM-matched within-platform DD. We note that other empirical papers have successfully combined CEM with difference-in-differences to identify causal effects (e.g., see Singh and Agrawal (2011)).

Thus far, we have examined two identification strategies: cross-platform DD, which relies on a perfectly matched Expedia control group, and within-platform DD, which relies on a CEM-matched TripAdvisor control group. The former strategy is robust against bias due to unobserved hotel differences, while the latter is robust against bias due to unobserved platform differences. As a final robustness check, we can put these two strategies together to identify the causal effect of management responses using a difference-in-difference-in-differences (DDD) design, which allows us to simultaneously control for cross-platform, and cross-hotel confounders. To implement DDD, we first need to identify a control group of hotels that should have been unaffected by treatment on either review platform. We again rely on the natural hotel matching available to us, and use all non-responding TripAdvisor hotels, and their corresponding 1-1 matched Expedia units. Conceptually, DDD takes place in two DD steps. First, we compute a cross-platform DD for affected hotels, similar to Equation 1. Then, we adjust this DD for unobserved cross-platform differences by subtracting from it the cross-platform DD for unaffected hotels. Formally stated, the DDD identification assumption is

$$E((Y_{i0}^{t+1} - Y_{i0}^{t}) - (Z_{i0}^{t+1} - Z_{i0}^{t})|D = 1, X) = E((Y_{i0}^{t+1} - Y_{i0}^{t}) - (Z_{i0}^{t+1} - Z_{i0}^{t})|D = 0, X).$$  (3)
3 Data

To study the effect of management review responses on hotel reputation we combine information collected from various sources. In this section, we describe the various datasets we collected, and then we explain how we merged them to obtain the sample we use in our analyses.

The two major sources of data we use are TripAdvisor and Expedia reviews for Texas hotels. TripAdvisor is a major travel review platform that contains more than 150 million reviews for millions of accommodations, restaurants, and attractions. TripAdvisor reached over 260 million consumers per month during 2013, a fact that signifies its influence on traveler decision making. We collected the entire review history of the 5,356 Texas hotels and accommodations that are listed on TripAdvisor. In total, our TripAdvisor sample contains 314,776 reviews, with the oldest review being from August 2001, and the most recent from December 2013. Each review in our dataset is associated with a star rating, text content, the date it was submitted, and a unique identifier for the reviewer who submitted it. If the review received a management response, we record the date the response was posted, which is typically differs from the date the review was submitted, and the content of the response. Out of the 5,356 hotels in our TripAdvisor sample, 4,603 received at least one review, and 2,590 left at least one management response.

Expedia is a web travel agency. Among the services it provides are airline and hotel reservations, and car rentals. Similar to TripAdvisor, consumers can review the Expedia services they purchase. We collected the entire review history of the 3,845 Texas hotels listed on Expedia, for a total of 519,962 reviews. The earliest Expedia review is from September 2004, and the most recent from December 2013. Our Expedia review sample contains the same review attributes as our TripAdvisor sample. Out of the 3,845 hotels in our Expedia sample, 3,356 were reviewed, and 587 left at least one management response.

Having collected TripAdvisor and Expedia reviews, our next step is to link these review samples together by hotel. To do so we exploit a feature of the Expedia website: Expedia
provides a link to each hotel’s TripAdvisor page, if such a page exists on TripAdvisor. This allows us to accurately match nearly every hotel’s Expedia and TripAdvisor reviews. To verify the accuracy of the Expedia provided link we randomly sampled 100 Expedia-TripAdvisor pairs, and manually verified that they correspond to the same hotel by checking the hotel’s name and address. We found no discrepancies. Using this information, we are able to match 3,681 out of 3,845 Expedia hotels (96% of the Expedia hotel sample). Of the 3,681 matched hotels 3,264 are reviewed on both sites. After matching each hotel across the two review platforms, we further balance our estimation sample by limiting ourselves to hotels that have been reviewed on both sites. This way, our data includes TripAdvisor and Expedia ratings for every hotel, and thus allows us to identify our treatment effect from only within-hotel, cross-platform variation. After limiting our sample to hotels that have been reviewed on both review platforms we are left with a total of 806,342 reviews out of which 291,119 are from TripAdvisor, and 515,223 from Expedia. Finally, since in our analyses we use Expedia ratings as a control group, we exclude from our data any hotels that have posted management responses on Expedia. This leaves us with 2,697 matched hotels, and 552,051 reviews of which 203,068 are from TripAdvisor, and 348,983 are from Expedia. Table 1 describes the various estimation samples we use in our analyses. The matched set of TripAdvisor and Expedia ratings for hotels that have been reviewed on both platforms, excluding hotels that have ever responded on Expedia constitutes our main estimation sample.4

**Construction the CEM-matched sample** One of the robustness checks we perform uses Coarsened Exact Matching (CEM) to strengthen the validity of the assumptions underlying within-platform DD estimation. We use CEM to construct an estimation sample where we match treated and control hotels on price category, operation, and city. We obtain

---

4We have conducted separate analyses with estimation samples that include the ratings of hotels that respond on Expedia up to the point they begin responding, as well as the ratings of hotels that have only been reviewed on one of the two review platforms. Our results are not sensitive to these alternative choices of estimation sample.
these hotel characteristics from Smith Travel Research (STR), a company which tracks hotel performance and maintains a census of all US hotels. The STR hotel census contains a rich set of attributes for all US hotels including price category (budget, economy, midprice, upscale or luxury), and operation (chain, franchise, or independent). We start by limiting our sample to the 4,603 hotels that are reviewed on TripAdvisor, reducing our dataset to 314,776 reviews. From these hotels, 2,590 responded to at least one reviews while the remaining 2,013 never did. Out of the 4,603 hotels, 1,468 are successfully matched by CEM, while no match is found for the remaining 3,135 hotels.

**Collecting user review histories** In Section 5, we use the entire TripAdvisor review history of every user who reviewed a Texas hotel on TripAdvisor. For every user that reviewed a hotel in our TripAdvisor sample, we collected his or her entire review history for a total of 3,047,428 reviews from 214,141 users. We were not able to obtain the review histories of a small fraction of users (2.2%) either because they left anonymous reviews on TripAdvisor (the username associated with such reviews is “A TripAdvisor Member”), or because they have closed their TripAdvisor accounts and therefore their user profiles do not exist anymore.

### 3.1 Descriptive statistics

A key difference between TripAdvisor and Expedia, which we exploit in our analysis, is that hotels often post management responses on TripAdvisor, but they rarely do so on Expedia. Figure 1 illustrates this difference: we plot the cumulative percentage of reviews that have received a management response by year. We find that by 2013, 31.5% of TripAdvisor reviews had received a management response compared to only 2.3% for Expedia, highlighting the difference in the rate of management response adoption across the two review platforms.

Having established that management responses are infrequent on Expedia, we next turn our attention to investigating the adoption patterns of management responses on TripAdvi-
An interesting aspect underlying the increasing adoption trend of management responses on TripAdvisor is the elapsed time between a review being posted and receiving a management response. Figure 2 plots the average lag (measured in days) between reviews and management responses by review submission year. On average, TripAdvisor reviews submitted in 2013 received a response 25 days later, while reviews posted in 2009 received a response almost 10 months later. How can we explain the managerial practice of responding to old reviews? A possible interpretation is that hotel managers are concerned that even old reviews can be read by, and affect the decision making process of future TripAdvisor visitors. By responding to these old reviews hotel managers are potentially attempting to steer the behavior of future TripAdvisor visitors who might stumble upon them.

Next we turn our attention to analyzing the frequency with which hotels respond to reviews on TripAdvisor. Figure 3 plots the fraction of TripAdvisor reviews that received a response by star-rating. While a priori we might expect negative reviews to be more likely to receive a response, we find that in our data this is not the case. In fact, 5-star reviews are among the most likely to receive a response, and negative reviews are almost as likely to receive a response as positive reviews.

What are the characteristics of hotels that use management responses? Table 2 compares hotels by their adoption of management responses on TripAdvisor. We find that responding hotels have higher average ratings both on TripAdvisor, and on Expedia. The mean difference between the star-ratings of responding and non-responding hotels is 0.5-stars. Table 2 also highlights an interesting cross-platform difference: while on average Texas hotels have more reviews on Expedia than they do on TripAdvisor, the length of the text associated with the average Expedia review is only one third of the length of the average TripAdvisor review. The average Expedia review is 201 characters long, only slightly longer than a tweet. This difference may further explain the reason behind the lower rate of adoption of management responses on Expedia: consumers do not write long, descriptive Expedia reviews that merit a response.
4 Results

In this section we present the results of regression analyses we carried out to estimate the causal effect of management responses on hotel reputation. These analyses are based on the three identification strategies we described above. In addition to these findings, we provide empirical evidence in support of the identification assumptions underlying our causal claims.

4.1 Cross-platform DD

Cross-platform DD, which is our preferred specification, estimates changes to the TripAdvisor ratings of any given hotel after it starts responding, relative to before, and adjusted for any change over the same period to its Expedia ratings. The identifying assumption that allows a causal interpretation of our findings is that TripAdvisor and Expedia ratings would have evolved in parallel in the absence of treatment. While this assumption isn’t fully testable, the panel nature of our data generates some testable hypotheses that we can use to reinforce the plausibility of our causal claims. Specifically, given our long observation period, we can test for differences in trends between the two platforms prior to treatment.

To compare pre-treatment trends, we partition time around around the day each hotel started responding in 30-day intervals, taking the offset of the first response to be 0. Then, for example, \([0, 30]\) is the 30-day interval starting on the day the hotel began responding, and \([-30, 0]\) is the 30-day interval just before. We focus our trend analysis on the two-year period centered on each hotel’s first response, resulting in the definition of 24 distinct intervals. Since hotels began responding at different times, these intervals correspond to different calendar dates for different hotels. Next, we associate each TripAdvisor and Expedia rating in our estimation sample with a dummy variable indicating the interval that contains it. Finally, we estimate the following DD regression

\[
\text{Stars}_{ijt} = \beta_1 \text{After}_{ijt} + \beta_2 \text{TripAdvisor}_j + \gamma \text{After}_{ijt} \times \text{TripAdvisor}_j \times \text{Interval}_{ijt} + \alpha_j + \tau_t + \epsilon_{ijt},
\]

(4)
where Stars$_{ijt}$ is the star-rating of review $i$ for hotel $j$ in calendar month $t$, After$_{ijt}$ is an indicator for reviews (on either platform) submitted after hotel $j$ started responding, TripAdvisor$_{ij}$ is an indicator for TripAdvisor ratings, and Interval$_{ijt}$ is the set of 30-day long treatment clock dummies we described above. The coefficient for After$_{ijt}$ captures differences in ratings between treatment and non-treatment periods, the coefficient for TripAdvisor$_{ij}$ captures differences in ratings across platforms, and $\gamma$, the vector of interaction coefficients associated with each interval, is the difference-in-differences estimate of interest. Finally, our model includes calendar-month fixed-effects $\tau_t$ to control for transient shocks in ratings that are common across review platforms.

While we could estimate this model by pooling ratings from different hotels together, we choose to include a matched-pair fixed effect $\alpha_j$, i.e., a shared fixed effect for reviews of the same hotel from either review platform. The use of matched-pair fixed effects enables identification from only within-hotel variation. In this statistical modeling choice, we follow the recent studies of Azoulay et al. (2013), Singh and Agrawal (2011), and Kovács and Sharkey (2014), who use the same technique when estimating matched DD models.

We estimate the model in Equation 4 using OLS. To account for serial correlation in our dependent variable, we cluster errors at the hotel level (Donald and Lang 2007, Bertrand et al. 2004). We choose to normalize the coefficient for the $[-60, -30)$ interval to 0. While choosing a different baseline would have yielded identical conclusions, our particular choice eases presentation as it will become evident shortly. The coefficients of the remaining intervals can be interpreted as differences between TripAdvisor and Expedia ratings over time with respect to the $[-60, 30]$ baseline. We present a graphical analysis of our estimates in Figure 4. The figure plots the estimated values of the interval coefficients $\gamma$, together with their 95% confidence intervals.

The figure reveals several distinctive features of hotel rating dynamics prior to, and following the adoption of management responses. First, visual inspection of pre-treatment

\footnote{The results of a pooled regression are not meaningfully different.}
trends suggests that they are parallel with the exception of the 30-day interval immediately preceding the treatment period. To back this claim statistically, we perform a Wald test, which fails to reject ($p < 0.43$) the hypothesis of joint equality among pre-treatment intervals excluding $[-30, 0)$.

Second, the figure reveals an outlier at $[-30, 0)$, which indicates that hotels adopt management responses as a reputation management strategy when they experience a substantive negative shock to their TripAdvisor ratings. This negative shock to TripAdvisor ratings prior to adopting management responses is reminiscent of Ashenfelter’s dip (Ashenfelter and Card 1984), an empirical regularity first observed in the context of job training programs, where program participants tended to experience an earnings drop just prior to enrolling in them. The presence of Ashenfelter’s dip can overstate our DD estimates because hotel ratings – just like employee earnings – are likely to mean revert following an out of the ordinary negative period regardless of any intervention by hotel management. Following common practice (see, e.g., Heckman and Smith (1999), Jepsen et al. (2014), Friedrich and Ording (2013), Li et al. (2011)), we correct for Ashenfelter’s dip by computing long-run differences, which symmetrically exclude a number of periods around the adoption of management responses. Our final observation regards the post-treatment period, and it foreshadows our main result. Following the adoption of management responses, we see a sustained increase in ratings. In fact, hotel ratings not only recover following the adoption of management responses, but they consistently exceed their prior levels by over 0.1 stars.

Given the graphical evidence in support of the parallel trends assumption underlying our identification strategy, we next estimate the causal impact of management responses on hotel ratings. The following model implements our cross-platform DD identification strategy:

$$
\text{Stars}_{ijt} = \beta_1 \text{After}_{ijt} + \beta_2 \text{TripAdvisor}_{ij} + \delta \text{After}_{ijt} \times \text{TripAdvisor}_{ij} + \alpha_j + \tau_t + \epsilon_{ijt}, \tag{5}
$$

The variables in this model are as in Equation 4, with the addition of controls $X_{ijt}$. As is common in DD analyses, we include review-platform specific quadratic time-trends in $X_{ijt}$ as
an additional safeguard against non-parallel trends. Again, the matched-hotel fixed effects \( \alpha_j \) ensure that our identification relies only on within hotel variation, \( i.e., \) comparing the ratings on any given hotel on TripAdvisor with the ratings of the same hotel on Expedia. The primary coefficient of interest is \( \delta \), which measures the causal impact of management responses on hotel ratings.

We first estimate Equation 5 on the sample of responding hotels using OLS with standard errors clustered at the hotel level. We present our results in the first column of Table 3. The estimated coefficient for the interaction term \( \text{After}_{ijt} \times \text{TripAdvisor}_{ij} \) is 0.15 stars, and it is highly statistically significant. Next, to correct for Ashenfelter’s dip, we repeat our estimation excluding ratings submitted anywhere between 30 days prior, and 30 days following a hotel’s first management response.\(^6\) We present these results in the second column of Table 3. As expected, our adjusted estimate for \( \delta \) is slightly smaller. However, even after accounting for transient negative shocks to hotel ratings prior to the response period, we find that management responses cause subsequent hotel ratings to rise by an average of 0.12 stars.

The coefficient for \( \text{After}_{ijt} \), which measures changes in the ratings of Expedia reviewers over the same time period is also of interest. We estimate its value to be statistically indistinguishable from zero, suggesting that Expedia reviewers were unaffected by management responses on TripAdvisor. This is as we would have hoped for, and provides additional evidence in support of the parallel trends identification assumption. If ratings for the control group had changed following treatment, it would be harder to argue that controlling for these changes completely eliminates bias. Moreover, the observation that the ratings of Expedia reviewers were unaffected by treatment indicates that it is highly unlikely that increased ratings after adopting management responses were the outcome of unobserved hotel improvements to avoid further negative reviews – unless one is willing to argue that only TripAdvisor reviewers experienced these improvements, and Expedia users did not see any

\(^6\)Sensitivity tests excluding longer periods did not yield meaningfully different results.
change whatsoever.

Overall our analysis suggests that responding hotels were able to significantly increase their future TripAdvisor ratings solely by responding to their past reviews. These findings indicate that management responses are a powerful reputation management tool that can improve consumer ratings and, in turn, financial performance.

4.2 Robustness checks using alternative identification assumptions

The cross-platform DD identification strategy is sensitive to bias arising from unobserved, platform-specific ratings shocks that are correlated with the adoption of management responses. For instance, something as simple as a user-interface change in the TripAdvisor review submission form could directly impact consumer ratings, leaving Expedia reviewers unaffected, and, consequently, rendering them a invalid control group. In this section, we seek to identify the causal impact of management responses on consumer ratings using alternative, but compatible with cross-platform DD, assumptions.

Our first robustness check, is based on the within-platform DD identification strategy outlined in Section 2.2. A natural test for whether the rise in ratings following the adoption of management responses is in fact caused by some unobserved platform-wide change on TripAdvisor is to compare the TripAdvisor ratings of responding, and non-responding hotels over the same time period. This identification strategy eliminates the concern of unobserved cross-platform variation in ratings biasing our results, and it provides for an alternative, independent test of the impact of management responses. We implement within-platform DD using the following model:

\[
\text{Stars}_{ijt} = \beta_1 \text{Responding}_j + \delta \text{After}_{ijt} \times \text{Responding}_j + X_{jt} + \eta_j + \tau_t + \epsilon_{ijt}
\]  

The variable \(\text{After}_{ijt} \times \text{Responding}_j\), whose coefficient is of interest, is an indicator for the ratings of responding hotels after they begin responding, \(\text{Responding}_j\) is an indicator for
the treatment group (i.e., hotels that eventually adopt management responses), \( \eta_j \) is a hotel fixed effect, and the remaining variables are as above. In the controls \( X_{jt} \) we include separate quadratic time-trends for responding and non-responding hotels, to further guard against the possibility of non-parallel outcomes for treated and control units. The primary difference between this specification, and cross-platform DD (Equation 1) is that in the absence of a natural 1-1 matching between treated and control units, differencing takes place with respect to calendar time (using the time fixed effects \( \tau_t \)) rather than with respect to the treatment clock.

We present the results of the within-platform DD specification in the first column of Table 4. The variable Responding\(_j\) is subsumed by the hotel fixed effects \( \eta_j \), and therefore its coefficient is not estimated. The DD estimate for After\(_{ijt} \times \text{Responding}_j \) is statistically significant, and its value is 0.13 stars \((p < 0.01)\), in line with our cross-platform DD analysis. Next, we correct for Ashenfelter’s dip by excluding a 60-day period centered on treatment. Our adjusted DD estimate is slightly smaller but of comparable magnitude (0.09 stars, \( p < 0.01 \)).

To further reduce endogeneity concerns arising from the absence of a natural 1-1 match between hotels, we re-estimate the above equation using the CEM-matched sample of hotels. The matching procedure guarantees that responding and non-responding hotels are similar along our chosen matching criteria, which are price, city, and operation. Out of the 4,603 TripAdvisor hotels with at least one review, 1,468 are successfully matched by CEM, and the remaining 3,135 are discarded, reducing the number of observations in our estimation sample by 68\%. We report the CEM-matched within-platform DD results, which incorporates a correction for Ashenfelter’s dip, in the second column of Table 4. Again, our estimates for the impact of management responses (0.08 stars, \( p < 0.01 \)) remain positive and significant.

As a final robustness check, we replicate our results using the DDD strategy, which is more stringent than the double differencing methods we have used thus far. Our estimation sample now comprises all responding and non-responding hotels on TripAdvisor, and their
1-1 matched control units on Expedia. Then, the DDD estimate compares post-treatment changes in TripAdvisor ratings for responding hotels against the baseline of matched Expedia ratings over the same period of time, and then adjusts this estimate for unobservable trends by differencing out cross-platform changes in ratings for non-responding hotels over the same period of time. In other words, the DDD estimator is the difference between the cross-platform DD for responding and non-responding hotels:

$$\text{DDD} = DD_{\text{cross-platform}}^{\text{responding}} - DD_{\text{cross-platform}}^{\text{non-responding}}$$

The following model implements our DDD estimator:

$$\text{Stars}_{ijt} = \beta_1 \text{Responding}_j + \beta_2 \text{TripAdvisor}_{ij} + \beta_3 \text{Responding}_j \times \text{TripAdvisor}_{ij} + \beta_4 \text{Responding}_j \times \tau_t + \beta_5 \text{TripAdvisor}_{ij} \times \tau_t + \delta_{\text{After}_{ijt}} \times \text{Responding}_j \times \text{Tripadvisor}_{ij} + X_{jt} + \alpha_j + \tau_t + \epsilon_{ijt} \tag{7}$$

The variables \(\text{Responding}_j \times \tau_t\), and \(\text{TripAdvisor}_{ij} \times \tau_t\) are a full-set of review-platform, and treatment status specific time fixed effects. The DDD estimate is \(\delta\). Because we can match TripAdvisor to Expedia ratings, we use matched-pair fixed effects \(\alpha_j\), which subsume the coefficient for \(\text{Responding}_j\). We report our results, first without and then with Ashenfelter’s dip correction, in Table 5. The DDD estimate for the impact of management responses on subsequent ratings, which controls for both cross-hotel and cross-platform unobservable trends as well as Ashenfelter’s dip, matches our results so far (0.11 stars, \(p < 0.01\)).

In summary, the consistency of the additional analyses we performed in this section with our main results increases our confidence in the causal interpretation of the treatment effect we estimate.
5 Why do management responses affect reputation?

In the previous section, we showed that when hotels start responding to their reviews they see a positive change to their consumer ratings that cannot be attributed to underlying quality investments. How can we explain the link between management responses and improved reputation? In this section, we draw upon two strands of research to provide a theoretical grounding for this finding. The first strand is the services marketing literature, and specifically studies of service failure and recovery. The second strand is studies of reciprocal behavior on reputation management platforms that allow bilateral feedback from both sides participating in a trade.

5.1 Management responses as a service recovery effort

Hotels regularly experience service failures, which have various undesirable consequences including consumer dissatisfaction, reduced consumer loyalty, and negative word of mouth (Lewis and McCann 2004). Prior research has shown that service recovery efforts can significantly improve consumers’ assessments of experiences that did not meet their expectations, and increase their likelihood of returning (Tax et al. 1998). While the literature identifies various service recovery strategies, the backbone of any recovery effort is a speedy apology that acknowledges the problem at hand, and outlines steps to rectify it (Smith et al. 1999). A hotel can use a management response to deliver such an apology, and to encourage a dissatisfied consumer to return and give the hotel a second chance. Such recovery efforts can result in increased consumer satisfaction, and potentially a fresh, positive review. In turn, positive reviews by these previously dissatisfied returning customers can improve a hotel’s reputation, and explain our findings. Here, we show that while management responses work as service recovery theory would predict, recovery efforts alone cannot explain the ratings increase hotels experience when they begin responding to consumers. Specifically, we show that reviewers who received a management response after leaving a negative review are more
likely to leave a second review than reviewers who did not receive a response. Moreover, this second review is on average more positive than their initial review. Nevertheless, while both of these findings are consistent with the predictions of service recovery theory, the number of reviews by returning consumers is so small that it cannot adequately explain the full extent of the ratings increase responding hotels experience.

In our data, 1.3% of TripAdvisor reviews are from consumers who have rated the same hotel in the past. Among responding hotels this fraction is slightly higher, 1.4%. We begin our analysis by testing the hypothesis that consumers who receive a management response are more likely to return, and leave a second review. To do so, we estimate the following logistic regression model

\[
\text{Returning consumer}_{kj} = \beta \text{Received response}_{kj} + \eta_j + \epsilon_{kj},
\]  

(8)

where Returning consumer\(_{kj}\) is an indicator variable that is set to 1 if consumer \(k\) has left more than one review for hotel \(j\), Received response\(_{kj}\) is an indicator variable set to 1 for consumers who received a management response for their initial review of hotel \(j\), and \(\eta_j\) is a hotel fixed effect. We present our results in the first column of Table 6. We find that consumers who received a management response are 9% more likely to provide a second review that those who didn’t receive response. Because hotels respond to positive as well as to negative reviews, and service recovery efforts are typically aimed at dissatisfied consumers, we repeat our analysis limiting our estimation sample to consumers whose initial review was below 3 stars. Our results, in the second column of Table 6, show that dissatisfied consumers who receive a response are even more likely \((43\% = e^{0.361})\) to return, consistent with the predictions of service recovery theory.

Next, we turn our attention to comparing the difference in ratings between a returning consumer’s first and second reviews as a function of receiving a management response for the first review.\(^7\) While we might expect any returning hotel guest to anticipate a better second review, it is likely that returning consumers who receive a management response will provide more positive feedback.

---

\(^7\)Even though consumers can write more than two reviews for the same hotel, few consumers in our data...
experience, we isolate the impact of management responses by estimating the additional change in ratings for consumers who received a management response compared against a baseline of returning consumers who didn’t. To do so, we construct a dataset containing the first and second reviews of every consumer who left at least two reviews, and we estimate the following regression model:

\[
\text{Stars}_{ikj} = \gamma_1 2^{nd \text{ review}}_{ikj} + \gamma_2 \text{Received response}_{kj} + \beta \text{Received response}_{kj} \times 2^{nd \text{ review}}_{ikj} + \eta_j + \epsilon_{ikj}.
\]

The dependent variable is the \(i^{th}\) rating of consumer \(k\) for hotel \(j\). Received response\(_{kj}\) is an indicator for consumers who received a response for their first review of hotel \(j\), and 2\(^{nd}\) review\(_{ikj}\) is an indicator for this being the consumer’s second review for hotel \(j\). As before, we limit our analysis to consumers whose first rating is below 3 stars. The coefficient of interest \(\beta\) has the standard DD interpretation. Our results, shown in Table 7, indicate that returning consumers are more positive by 0.8 stars, but those who receive a response increase their second ratings by almost half a star more, highlighting the effect of recovery efforts. Unfortunately, as indicated by the small sample size of this analysis (\(N = 400\)), the aggregate effect of such improved ratings on hotel reputation is insignificant. In fact, our main results from Section 4 remain practically unchanged when we exclude returning reviewers from our data.\(^8\). Therefore, while management responses can contribute to the recovery of individual consumers who experienced a service failure, the total number of reviews created from such recovery efforts is too small to adequately explain the magnitude of the effect of management responses on hotel ratings.

---

\(^8\) For brevity, we do not show these regressions, as their results are nearly identical to those in Tables 3, 4, and 5.
5.2 Management responses as reciprocal reviews

An alternative way management responses could affect hotel reputation derives from their public nature. While management responses are directed to, and displayed below specific reviews, future TripAdvisor visitors can also read them alongside a hotel’s reviews. How can management responses to past hotel guests increase a hotel’s future ratings? Here, we argue that management responses are a signaling mechanism that shifts reviewer selection towards reviewers that are inherently more positive in their evaluations.

We base the hypothesis that management responses can affect reviewer selection on the literature in reciprocity on review platforms. A number of field and lab studies (Resnick and Zeckhauser 2002, Dellarocas and Wood 2008, Bolton et al. 2013) have established that in settings where agents can sequentially rate each other there is a dearth of negative ratings, which arises from fear of reciprocation. The primary example of this phenomenon is eBay. Up to 2008, during which time eBay buyers and sellers could rate each other, buyers with a poor experience would often avoid leaving a negative review for a seller for fear that the seller would also follow up with negative review. When eBay introduced new rules that removed the option for sellers to leave negative feedback for buyers, sellers started receiving an increased number of negative reviews (Hui et al. 2014).

Here, we draw a parallel between eBay and TripAdvisor: we think of responding hotels as participating in a review platform where reciprocation is a possibility – similar to eBay’s old reputation system – and, non-responding hotels as participating in a review platform where reviews will not be reciprocated – similar to eBay’s new reputation system. By responding to past guests, hotels signal to future reviewers that they may reciprocate a bad review, discouraging some guests with a negative experience from leaving a review altogether. Overall, this behavior can shift reviewer selection towards reviewers with higher ratings, and, on average, improve the ratings a responding hotel receives.

The reciprocation theory generates a testable hypothesis for our data that is based on the observation that a reviewer’s rating for a specific business can be thought of as consisting of
three separable components: the reviewer’s type, \textit{i.e.}, how a positive, or negative a reviewer is overall; the business’s type, \textit{i.e.}, the overall quality of a business; and, the reviewer’s experience with a specific business.\textsuperscript{9} Concretely, we think of the rating of reviewer \( k \) for business \( j \) as

\[
\text{Stars}_{jk} = \theta_k + \eta_j + \epsilon_{jk}, \tag{10}
\]

and the reviewer’s decision to submit a review depends on \( \text{Stars}_{jk} \) exceeding a unobserved user-specific threshold.\textsuperscript{10}

Then, the reciprocation theory suggests that consumers with a lower type (\textit{i.e.}, lower values of \( \theta_k \)) will be less likely to leave a review for responding hotels. To operationalize a reviewer’s type, we estimate the reviewer fixed effects \( \theta_k \) using the sample of each reviewer’s entire review history. Then, we hypothesize that the distribution of reviewer types will be more positive following a hotel’s adoption of management responses. To test this, we estimate the following regression model on the sample of responding and non-responding TripAdvisor hotels

\[
\text{Reviewer type}_{kjt} = \beta \text{After}_{kjt} + \eta_j + \tau_t + \epsilon_{ijt}, \tag{11}
\]

where the dependent variable \( \text{Reviewer type}_{kjt} \) is the value of \( \theta_k \) associated with the consumer \( k \) who reviewed hotel \( j \) at time \( t \) as estimated from Equation 10. \( \text{After}_{ijt} \) is an indicator for reviewers who submit reviews after hotel \( j \) starts responding. The coefficient of interest, \( \beta \), captures the change in the inherent positivity of reviewers of responding hotels after their start responding. To further limit the influence of unobserved transient factors that could affect reviewer selection, we borrow an idea from regression discontinuity designs: we limit our estimation sample to 1 year before, and 1 year after the treatment, since any two reviewers are more likely to be comparable in their unobserved characteristics if their

\textsuperscript{9}Dai et al. (2012) take a similar approach in deconstructing consumer ratings, and demonstrate how it provides a more accurate prediction of a business’ true quality.

\textsuperscript{10}This simple model can easily be extended to accommodate users with low ratings who wouldn’t have otherwise submitted a review, and who decide to artificially inflate their ratings and leave a review.
reviews are closer in time. We present our results in Table 8. As hypothesized, we find that management responses affect reviewer selection: reviewers who submit reviews after hotels start responding are 0.09 stars more positive than the population of reviewers who submitted reviews prior to management responses. A robustness check with smaller bandwidth of 6 months yields similar results.

How strong is the link between better ratings following the adoption of management responses and a change in reviewer selection? If the sole effect of management responses were to attract consumers who leave reviews that are on average 0.09-stars higher, we should expect to see a 0.09-star increase in the ratings of responding hotels. This estimate is close in magnitude to the increase in ratings hotels see when they adopt management responses. While with the data available to us we cannot rule out the possibility of a more complex set of processes to explain the impact of management responses, the similarity of these two effects, which were estimated on distinct datasets, suggests that change in reviewer selection is an influential component underlying the reputation improvement of responding hotels.

Overall, the reciprocation theory is appealing because it can explain both the difference in ratings between responding and non-responding hotels on TripAdvisor, as well as the difference between a responding hotel’s ratings on TripAdvisor and Expedia. However, we recognize that its application to our setting has limitations. Hotels cannot provide an actual star-rating for their guests, which would visibly affect their online reputation. Moreover, it is unclear what economic consequences an argumentative management response could have for a reviewer that would affect the reviewer’s strategic considerations in leaving a negative review in the first place. Nevertheless, existing research (Ockenfels et al. 2012) suggests that consumers place more value on their online reputation than economic incentives alone would predict. Therefore, while there does not appear to be an obvious economic cost to receiving an argumentative management response, there are social and emotional costs that can affect a reviewer’s decision to leave a negative review. In addition, the observations that hotels respond to five-star reviews as often as one-star reviews, as well as to years-old
reviews lend more credence to the argument that responding is more about signaling than it is about service recovery. Given the impact of management responses on consumer behavior, developing better theoretical models of how consumers evaluate a firm’s responses to their reviews is a promising direction for future research.

6 Discussion and conclusions

In this paper, we show that management responses are an effective way for firms to improve their online reputation. We study the Texas hotel industry, and we show that, on average, responding hotels see a 0.12-star increase in their TripAdvisor ratings when they begin responding to reviewers. To explain this finding, we consider the possibility that improved reputation is the outcome of unobserved quality investments or service recovery efforts, but we do not find support for these theories. Instead, we hypothesize that management responses cause an underreporting of negative reviews. We base this hypothesis on prior research (e.g., Bolton et al. (2013)) that has shown that negative reviews are rarer on review platforms that allow bilateral feedback. We then argue that the transition from not responding to consumer reviews to responding to them is analogous to a transition from a unilateral feedback system to bilateral one. We empirically support this hypothesis by showing that inherently negative reviewers become less likely to leave a review for responding hotels.

Our findings have economic and managerial implications for hotels, consumers, and review platforms. As far as hotels are concerned, our results indicate that management responses are an effective reputation management strategy. Further, this strategy is sanctioned by review platforms, and it can directly impact the financial performance of firms that use it (Luca 2011). The benefits of management responses for consumers and review platforms are less obvious. On one hand, by opening up a communication channel to consumers, review platforms encourage hotels to engage with their guests, to inform future visitors of steps they
have taken to correct issues reported in prior reviews, and to create a richer information environment that should in principle help consumers make better choices. On the other hand, our work shows that opening up this communication channel has the undesired consequence of negative review underreporting, which creates a positive bias in the ratings of responding hotels.

Taken together, these observations highlight an information design problem: how can review platforms enable the interaction of firms and consumers without introducing reporting bias? While it is beyond the scope of this work to provide an exhaustive list of alternative designs, other practical schemes to consider include responding to consumers privately, and management responses that are not attached to specific reviews. The evaluation and implementation of these and other information design schemes is a direction for future research.

A potential limitation of our analysis is that we focus on estimating a treatment effect on the treated. In other words, our work does not answer the question of what the impact of using management responses would be on the reputation of randomly chosen hotel. In practice, the treatment effect could be significantly different for hotels that do not currently respond to their reviews. We speculate that this is unlikely to be the case given that our analysis indicates that the primary driver of improved reputation is a change in reviewer behavior rather than any particular hotel characteristic. Nevertheless, we caution that the possibility remains that management responses could affect the reviewing habits of the guests of non-responding hotels in different ways. Therefore, our findings should not be interpreted as a definite prescription for improving a firm’s online reputation, but rather as a promising reputation management strategy that is worth investigating. A randomized field experiment is a promising path forward in estimating any difference between the ATT and the ATE.

A second limitation of our work is that we solely focus on the decision of hotel managers to begin responding to consumer reviews. Managers who are considering responding to consumer reviews are facing a complex decision problem that involves choosing which reviews to respond to, when to respond to them, and how to respond. Future work can combine
econometric methods with natural language processing to analyze the textual content of reviews and management responses to estimate heterogeneity in the treatment effect arising from the various ways businesses handle praise and complaints. Such analyses can yield prescriptive guidelines for managers looking to communicate with consumers in different customer service scenarios.

Online reviews have been thoroughly studied in the marketing, management, and computer science literatures, and they are now understood to be a significant driver of consumer behavior. By comparison, existing research has overlooked management responses despite their wide adoption by managers. In this paper, we took a first step in narrowing this gap by showing that management responses can shape online reputation. We also found that more than half of the hotels we studied now respond to their reviews, up from only 7% five years ago. Given the increasing number of firms that wish to engage with their reviewers and to actively participate in the production of information that molds their online reputation, we look forward to seeing more research in this direction.
References

Ashenfelter, Orley C, David Card. 1984. Using the longitudinal structure of earnings to estimate the effect of training programs.


Friedrich, Timo, Magnus Ljungberg Ording. 2013. Effects of job-training programs on earnings.


Hui, Xiang, Maryam Saeedi, Zeqian Shen, Neel Sundaresan. 2014. From lemon markets to managed markets: the evolution of ebays reputation system.


Jepsen, Christopher, Kenneth Troske, Paul Coomes. 2014. The labor-market returns to community college degrees, diplomas, and certificates. *Journal of Labor Economics* 32(1) 95–121.

Kovács, Balázs, Amanda J Sharkey. 2014. The paradox of publicity: How awards can negatively affect the evaluation of quality. *Administrative Science Quarterly* 59(1) 1–33.


Figure 1: The cumulative number of responses over the cumulative number of reviews posted by year.
Figure 2: Average lag (in days) between a TripAdvisor review and its management response by review submission year. We omit years prior to 2009 because of insufficient observations.
Figure 3: The fraction of TripAdvisor reviews that received a management response by star-rating. The dashed line is the overall average which is 31.5%.
Figure 4: The evolution of treatment effects, i.e., differences in hotel ratings between Expedia and TripAdvisor, as a function of a hotel’s decision to begin responding to reviews. The solid line plots the $\gamma$-coefficient estimates from Equation 4, and the dashed lines their respective 95% confidence intervals.
Table 1: Dataset description.

<table>
<thead>
<tr>
<th></th>
<th>TripAdvisor</th>
<th>Expedia</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>All hotels</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reviewed</td>
<td>4,603</td>
<td>3,356</td>
</tr>
<tr>
<td>Responding</td>
<td>2,590</td>
<td>587</td>
</tr>
<tr>
<td>Reviews</td>
<td>314,776</td>
<td>519,962</td>
</tr>
<tr>
<td>Responses</td>
<td>99,178</td>
<td>11,781</td>
</tr>
<tr>
<td><strong>Matched hotels</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reviewed</td>
<td>3,511</td>
<td>3,265</td>
</tr>
<tr>
<td>Responding</td>
<td>2,387</td>
<td>568</td>
</tr>
<tr>
<td>Reviews</td>
<td>296,138</td>
<td>515,227</td>
</tr>
<tr>
<td>Responses</td>
<td>97,684</td>
<td>11,779</td>
</tr>
<tr>
<td><strong>Matched hotels reviewed on both platforms</strong></td>
<td>3,264</td>
<td>3,264</td>
</tr>
<tr>
<td>Responding</td>
<td>2,303</td>
<td>567</td>
</tr>
<tr>
<td>Reviews</td>
<td>291,119</td>
<td>515,223</td>
</tr>
<tr>
<td>Responses</td>
<td>96,665</td>
<td>11,776</td>
</tr>
<tr>
<td><strong>Cross-platform DD hotels</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reviewed</td>
<td>166,152</td>
<td>263,804</td>
</tr>
<tr>
<td>Responses</td>
<td>55,684</td>
<td>−</td>
</tr>
<tr>
<td><strong>Within-platform DD hotels</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reviewed</td>
<td>314,776</td>
<td>−</td>
</tr>
<tr>
<td>Responses</td>
<td>99,178</td>
<td>−</td>
</tr>
<tr>
<td><strong>Cross-platform DDD hotels</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reviewed</td>
<td>203,068</td>
<td>348,983</td>
</tr>
<tr>
<td>Responses</td>
<td>55,684</td>
<td>−</td>
</tr>
</tbody>
</table>

\[11\] Matched responding hotels that are reviewed on both platforms, excluding hotels that respond on Expedia

\[12\] Hotels that are reviewed on TripAdvisor

\[13\] Matched hotels that are reviewed on both platforms, excluding hotels that respond on Expedia
Table 2: Hotel summary statistics.

<table>
<thead>
<tr>
<th></th>
<th>TripAdvisor</th>
<th>Expedia</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Matched Hotels</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Avg. hotel rating</td>
<td>3.6</td>
<td>3.9</td>
</tr>
<tr>
<td>Reviews per hotel</td>
<td>84.3</td>
<td>157.8</td>
</tr>
<tr>
<td>Responses per hotel</td>
<td>27.8</td>
<td>3.6</td>
</tr>
<tr>
<td>Avg. review length</td>
<td>617.0</td>
<td>201.0</td>
</tr>
<tr>
<td>Avg. response length</td>
<td>439.2</td>
<td>306.5</td>
</tr>
<tr>
<td><strong>Matched hotels that respond on TripAdvisor</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Avg. hotel rating</td>
<td>3.8</td>
<td>4.1</td>
</tr>
<tr>
<td>Reviews per hotel</td>
<td>107.4</td>
<td>183.7</td>
</tr>
<tr>
<td>Responses per hotel</td>
<td>40.9</td>
<td>5.0</td>
</tr>
<tr>
<td>Avg. review length</td>
<td>624.3</td>
<td>200.2</td>
</tr>
<tr>
<td>Avg. response length</td>
<td>439.2</td>
<td>307.2</td>
</tr>
<tr>
<td><strong>Matched hotels that don’t respond on TripAdvisor</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Avg. hotel rating</td>
<td>3.3</td>
<td>3.6</td>
</tr>
<tr>
<td>Reviews per hotel</td>
<td>35.4</td>
<td>95.7</td>
</tr>
<tr>
<td>Responses per hotel</td>
<td>—</td>
<td>0.2</td>
</tr>
<tr>
<td>Avg. review length</td>
<td>601.3</td>
<td>203.0</td>
</tr>
<tr>
<td>Avg. response length</td>
<td>—</td>
<td>291.6</td>
</tr>
</tbody>
</table>

Table 3: Cross-platform DD.

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>After × Tripadvisor</td>
<td>0.145***</td>
<td>0.119***</td>
</tr>
<tr>
<td></td>
<td>(6.72)</td>
<td>(5.05)</td>
</tr>
<tr>
<td>Tripadvisor</td>
<td>−0.790***</td>
<td>−0.802***</td>
</tr>
<tr>
<td></td>
<td>(−4.52)</td>
<td>(−4.55)</td>
</tr>
<tr>
<td>After</td>
<td>−0.003</td>
<td>−0.009</td>
</tr>
<tr>
<td></td>
<td>(−0.28)</td>
<td>(−0.72)</td>
</tr>
<tr>
<td>Ashenfelter’s dip correction</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>N</td>
<td>429956</td>
<td>415157</td>
</tr>
<tr>
<td>R² within</td>
<td>0.020</td>
<td>0.020</td>
</tr>
</tbody>
</table>

Note: The dependent variable is the review i star rate of hotel j at time t. Cluster-robust t-statistics (at the individual hotel level) are shown in parentheses. All specifications include time fixed effects and platform specific quadratic time trends. Significance levels: * p<0.1, ** p<0.05, *** p<0.01.
Table 4: Within platform DD.

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>After × Responding</td>
<td>0.125***</td>
<td>0.087***</td>
<td>0.080***</td>
</tr>
<tr>
<td></td>
<td>(8.85)</td>
<td>(5.54)</td>
<td>(3.07)</td>
</tr>
<tr>
<td>Ashenfelter’s dip correction</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>CEM Sample</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>N</td>
<td>314776</td>
<td>256686</td>
<td>94442</td>
</tr>
<tr>
<td>R² within</td>
<td>0.0098</td>
<td>0.011</td>
<td>0.012</td>
</tr>
</tbody>
</table>

Note: The dependent variable is the review $i$ star rate of hotel $j$ at time $t$. Cluster-robust t-statistics (at the individual hotel level) are shown in parentheses. All specifications include time fixed effects.

Significance levels: * $p<0.1$, ** $p<0.05$, *** $p<0.01$.

Table 5: DDD.

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>After × Responding × Tripadvisor</td>
<td>0.113***</td>
<td>0.108***</td>
</tr>
<tr>
<td></td>
<td>(6.59)</td>
<td>(5.47)</td>
</tr>
<tr>
<td>Ashenfelter’s dip correction</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>N</td>
<td>552051</td>
<td>415157</td>
</tr>
<tr>
<td>R² within</td>
<td>0.021</td>
<td>0.020</td>
</tr>
</tbody>
</table>

Note: The dependent variable is the review $i$ star rate of hotel $j$ at time $t$. Cluster-robust t-statistics (at the individual hotel level) are shown in parentheses. All specifications include time fixed effects and platform specific quadratic time trends.

Significance levels: * $p<0.1$, ** $p<0.05$, *** $p<0.01$.

Table 6: Logistic regression of the probability of a second review by the same consumer as a function of receiving a management response.

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>All Reviews</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Received response</td>
<td>0.088**</td>
<td>0.361**</td>
</tr>
<tr>
<td></td>
<td>(1.99)</td>
<td>(2.10)</td>
</tr>
<tr>
<td>Only stars &lt; 3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>211424</td>
<td>7023</td>
</tr>
</tbody>
</table>

Note: The dependent variable is an indicator of whether a consumer reviewed the same hotel twice. The independent variable is an indicator of whether a consumer’s first review received a response. All the specifications include hotel fixed effects.

Significance levels: * $p<0.1$, ** $p<0.05$, *** $p<0.01$. 
Table 7: Change in the star rating of a consumer’s second review, as function of the first review receiving a response.

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Received Response ×</td>
<td></td>
</tr>
<tr>
<td>Second Review</td>
<td>0.461**</td>
</tr>
<tr>
<td></td>
<td>(2.17)</td>
</tr>
<tr>
<td>Received Response</td>
<td>0.363</td>
</tr>
<tr>
<td></td>
<td>(0.56)</td>
</tr>
<tr>
<td>Second Review</td>
<td>0.832***</td>
</tr>
<tr>
<td></td>
<td>(7.06)</td>
</tr>
</tbody>
</table>

N: 400  
R² within: 0.32

Note: The dependent variable is the star of the current review.  
Significance levels: * p<0.1, ** p<0.05, *** p<0.01.

Table 8: Change in reviewer types following a hotel’s decision to begin responding.

<table>
<thead>
<tr>
<th></th>
<th>(1) BW= ±6 months</th>
<th>(2) BW= ±12 months</th>
</tr>
</thead>
<tbody>
<tr>
<td>After</td>
<td>0.094***</td>
<td>0.090***</td>
</tr>
<tr>
<td></td>
<td>(8.65)</td>
<td>(10.05)</td>
</tr>
</tbody>
</table>

N: 42600  
R² within: 0.0027

Note: The dependent variable is the reviewer type $\theta_k$ associated with the consumer $k$ who reviewed hotel $j$ at time $t$. Cluster-robust t-statistics (at the individual hotel level) are shown in parentheses. All specifications include hotel fixed effects.  
Significance levels: * p<0.1, ** p<0.05, *** p<0.01.