

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

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

Georgios Zervas
Questrom School of Business
Boston University
zg@bu.edu

November 7, 2016

Abstract

We investigate the relationship between a firm's use of management responses and its online reputation. We focus on the hotel industry, and present several findings. First, hotels are likely to start responding following a negative shock to their ratings. Second, hotels respond to positive, negative, and neutral reviews at roughly the same rate. Third, by exploiting variation in the rate with which hotels respond on different review platforms and variation in the likelihood with which consumers are exposed to management responses, we find a 0.12-star increase in ratings and a 12% increase in review volume for responding hotels. Interestingly, when hotels start responding they receive fewer but longer negative reviews. To explain this finding, we argue that unsatisfied consumers become less likely to leave short indefensible reviews when hotels are likely to scrutinize them. Our results highlight an interesting trade-off for managers considering responding: fewer negative ratings at the cost of longer and more detailed negative feedback.

[†]We thank Frederic Brunel, John Byers, Sharon Goldberg, Michael Luca, 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 experience failures that lead to negative reviews. Dealing with negative reviews is challenging because, unlike offline word-of-mouth, they persist online and firms can neither selectively delete them, nor opt-out from being reviewed altogether. To manage unfavorable reviews, firms often resort to questionable practices like review fraud (Mayzlin et al., 2014; Luca and Zervas, 2015), soliciting positive reviews in exchange for perks, threatening legal action against negative reviewers, and using non-disparagement clauses in sales contracts that stipulate fines if consumers write negative reviews. At the same time, technological advances in detecting fake 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 is 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 affect the responding firm's average rating. While review platforms ensure that responses meet 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. Yet, despite management responses now being commonplace, their efficacy in recovering a firm's reputation remains an open question.

In this paper, we estimate the impact of management responses on TripAdvisor hotel ratings. We show that, on average, responding hotels see a consistent increase of 0.12 stars in their ratings after they start using management responses. While this gain appears modest when evaluated against the usual 5-star scale, in practice, most ratings are concentrated to a narrower range. The standard deviation of hotel ratings in our data is 0.8 stars. Further, because TripAdvisor and other review platforms round average ratings to the nearest half-star, small changes 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. In our data, 27% of responding hotels increased their rounded ratings by at least half a star within 6 months of their first management response.

Several selection issues need to be considered before ascribing a causal interpretation to our results. First, hotels select into treatment, *i.e.*, responding to reviews. Second, hotels choose which reviews to respond to and how to respond to them. If unaccounted for, these non-random choices can bias estimation of an *average treatment effect* (ATE). For instance, our estimate could be biased upwards if we do not account for the possibility that hotels that are "better" at responding

are also more likely to respond. Convincingly controlling for these choices is difficult outside an experimental context.

Thus, instead of estimating an ATE, our goal is to consistently estimate an *average treatment effect on the treated* (ATT). The ATT can be consistently estimated when treatment assignment is non-random, and in particular when there is correlation between treatment and potential outcomes, *e.g.*, if hotels decide to respond based on an expectation that responding will increase their ratings. The ATT measures the impact of management responses conditional on the hotels that self-selected into treatment, the reviews they decided to respond to, and the manner in which they responded. The ATT will be biased if a hotel’s decision to respond is driven by unobserved factors that also affect the hotel’s ratings. For instance, a hotel’s decision to respond may be prompted by (unobserved to us) service improvements and renovations that the hotel made to avoid further negative reviews.¹ 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 approach this identification challenge in various ways requiring different assumptions from the data. Table 3 summarizes our identification strategies and robustness checks, which we describe in detail next.

Our first identification strategy uses Expedia ratings to control for changes in hotel quality. This approach is motivated by a difference in managerial practice between TripAdvisor and Expedia: while hotels frequently respond to TripAdvisor reviews, they almost never do so on Expedia. We build upon this observation to estimate an ATT using a difference-in-differences (DD) identification strategy. Intuitively, the DD estimator compares 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. The key assumption needed for the DD estimate to be consistent is that differences between TripAdvisor and Expedia ratings would have been constant in the absence of treatment. To defend this assumption, we need to understand why hotels respond on one platform but not the other.

Is the choice to only respond on TripAdvisor exogenously determined, or is it driven by changes in hotel quality? One explanation for solely responding on TripAdvisor that is compatible with our identification assumptions is that reviews are less salient on Expedia. Unlike TripAdvisor, which is in the business of collecting and disseminating reviews, Expedia is an online travel agency (Mayzlin et al. (2014) make the same point.) Comparing how the two sites present information 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 and responses are only available on a secondary page. In addition to being displayed less prominently, Expedia reviews are much shorter, and they arrive at nearly twice the rate they do on TripAdvisor. Therefore, hotels may be less inclined to respond to them because they are less substantive and are quickly superseded by fresher information. Another motivation for hotels to respond more

¹A recent *New York Times* article suggests that hotels commonly use online reviews as a guide for renovations. See <http://www.nytimes.com/2014/09/23/business/hotels-use-online-reviews-as-blueprint-for-renovations.html>.

frequently on TripAdvisor is that, unlike Expedia, TripAdvisor allows non-verified hotels guests to submits reviews. Therefore, hotels may be more likely to closely monitor TripAdvisor and respond to negative reviews they perceive as unfair or fake.

Cross-platform DD will be biased if hotels take other actions that affect their TripAdvisor ratings relative to Expedia at the same time they start responding. For instance, if hotels make renovations specifically valued by TripAdvisor users, which they then announce by responding to TripAdvisor reviews, the ATT we estimate will be likely biased upwards. We perform several robustness checks to show that our results are unlikely to be driven by TripAdvisor-specific improvements. First, we show that for a long period preceding each hotel’s first management response, TripAdvisor and Expedia ratings moved in parallel. Therefore, at least prior to treatment, ratings on the two review platforms are consistent with TripAdvisor and Expedia users valuing changes in hotel quality equally. Second, we show that management responses on TripAdvisor had no impact on the same hotel’s Expedia ratings. Therefore, for our estimate to be biased it would have to be the case that Expedia users have no value whatsoever for hotel improvements targeted at TripAdvisor users. Third, consider the possibility that hotels make TripAdvisor-specific improvements by targeting a traveler segment that is over-represented on TripAdvisor compared to Expedia. For example, if business travelers strongly prefer TripAdvisor and hotels make improvements specifically valued by business travelers, TripAdvisor ratings will rise relative to Expedia. We argue that this is unlikely to be the case because our results hold even when we compare TripAdvisor and Expedia travelers belonging to the same segments. Fourth, we show that the impact of management responses is larger for reviewers that are more likely to have read them. A reviewer’s propensity to read management responses is outside a hotel’s control, and is therefore unlikely to be correlated with unobserved actions the hotel took to improve its ratings.

A related concern arises if hotels simultaneously adopt multiple reputation management strategies. For instance, some hotels may start posting fake reviews at the same time they start responding (Mayzlin et al., 2014; Luca and Zervas, 2015). This is particularly problematic in our setting because posting fake reviews is easier on TripAdvisor than it is on Expedia. To ensure that the ATT we estimate is not driven by review fraud, we show that our results hold for hotels that are unlikely to commit review fraud in the first place.

To avoid bias due to cross-platform differences, we develop a second identification strategy that only relies on TripAdvisor ratings. The basic idea behind this strategy is that any difference in the ratings of two guests who stayed at the same hotel at the same time is unlikely to be due to unobserved hotel improvements. Thus, we estimate the impact of management responses by comparing the ratings of guests who left a review before a hotel began responding with the ratings of guests who stayed at the same hotel at the same time but left a review after the hotel began responding. This estimate is nearly identical to our cross-platform estimate.

In the last part of the paper, we turn our attention to understanding the mechanism underlying our findings. We argue that management responses result in better ratings because they change the cost of leaving a review in two ways. First, we argue that management responses decrease the

cost of leaving a positive review because consumers have positive utility for hotel managers taking notice of their feedback. Conversely stated, consumers may choose not to leave a positive review, if they are unsure hotel managers will read it. Second, we argue that management responses increase the cost of leaving a negative review because reviewers know that their feedback will be scrutinized.

We provide evidence for this mechanism by investigating the impact of management responses on two additional outcomes managers care about: review volume and review length. First, we examine the argument that consumers are more willing to leave a review if managers are likely to notice their feedback. To do this, we show that review volume increases following the adoption of management responses. Further, we show that after hotels start responding, they attract more reviewers who are more positive in their evaluations even when they review non-responding businesses, suggesting that these positive reviewers see management responses as an incentive to leave a review. Next, we examine the argument that management responses increase the cost of leaving a negative review. We show that, when hotels respond, even though negative reviews become more infrequent, they also become longer. Meanwhile, the length of positive reviews remains the same. This suggests that when hotel guests have a poor experience they may opt out of leaving a review unless they are willing to invest the extra effort required to write a defensible complaint. While some reviewers will choose to expend this extra effort, others will not. Thus, when hotels start responding they attract fewer but longer negative reviews. On one hand, these longer negative reviews may alarm hotel managers considering responding. On the other, however, they are in fact a natural side-effect of the mechanism driving the overall increase in positive ratings. This highlights an interesting trade-off in using management responses: better ratings at the cost of fewer but longer negative reviews.

2 Empirical strategy

Our goal is to estimate the 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 of TripAdvisor modified to omit management responses, but is identical otherwise. Then, using counterfactual notation, for any responding hotel i the ATT is given by:

$$E(Y_{i1} - Y_{i0} | D = 1)$$

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 arising from 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. 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} .

Before describing our identification strategy for the ATT, we highlight some difficulties inherent in estimating an ATE even with a randomized controlled trial. Unlike the hypothetical ATT experiment that randomly exposes some users to management responses, to estimate an ATE, we would have to instruct a randomly chosen set of hotels to start responding. We would also have to instruct these hotels on which reviews to respond to. While this could also be done at random, it is hard to argue that this strategy is close to what hotels might do in practice. Next, we would next have to randomize the types of responses treated hotels post. For example, should hotels respond in an antagonistic or in a conciliatory manner? In depth, or briefly? The space of treatments (*i.e.*, response strategies) seems so large that, unless we want to estimate the ATE of a specific strategy, focusing on the impact of management responses given the way hotels currently respond (*i.e.*, the ATT) seems more sensible.

2.1 Cross-platform 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 $E(Y_{i0}^{pre}|D = 1) = E(Y_{i0}^{post}|D = 1)$.² This assumption 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) identi-

²For 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.

fication 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

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

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. DD allows both for platform-independent transient shocks to hotel ratings as well as time-invariant cross-platform differences in hotel ratings. We can partially test the parallel-trends assumption by comparing the pre-treatment rating 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.

Triple differences As a robustness check, we also estimate the 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 responding hotels, similar to Equation 1. Then, we adjust this DD for unobserved cross-platform differences by subtracting from it the cross-platform DD for non-responding hotels. Formally stated, the DDD identification assumption is

$$E((Y_{i0}^{t+1} - Y_{i0}^t) - (Z_{i0}^{t+1} - Z_{i0}^t) | D = 1, X) = E((Y_{i0}^{t+1} - Y_{i0}^t) - (Z_{i0}^{t+1} - Z_{i0}^t) | D = 0, X). \quad (2)$$

2.2 Within-platform identification strategy

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. In this section, we describe an identification strategy to mitigate this concern. Our approach exploits the fact that most (over 98%) TripAdvisor reviewers indicate in their reviews when they stayed in a hotel. The insight motivating this identification strategy is that any difference in the ratings of two TripAdvisor reviewers who stayed at the same hotel at the same time is unlikely to be driven by unobserved hotel renovations. This model only relies on variation in the ratings of guests who stayed at the same hotel in the same month to identify the impact of management responses.

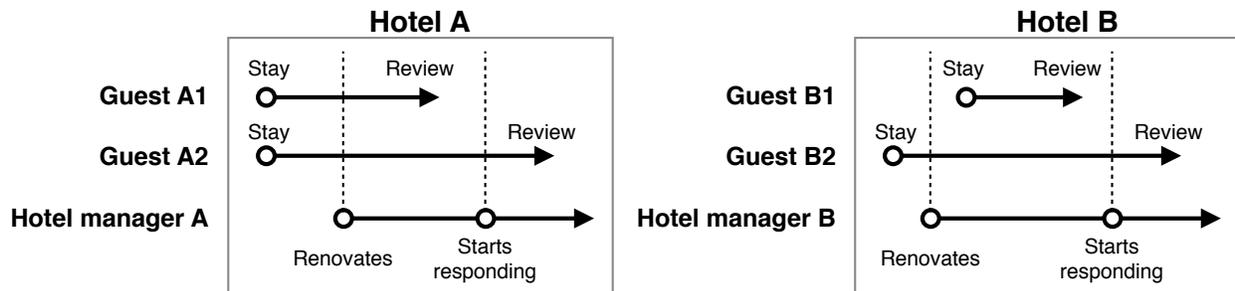


Figure 1: Within-platform identification relies on the reviews of hotel A but not hotel B.

Figure 1 illustrates how this identification strategy solves the problem of unobserved hotel renovations. Within-platform identification of the impact of management responses conditional on guests' dates of stay relies on the difference between reviews A1 and A2 but not B1 and B2. Hotel A's unobserved renovation is not a concern because guests A1 and A2 stayed at the hotel at the same time. By contrast, a comparison of reviews B1 and B2 could result in bias when estimating the impact of management responses because guest B2 experienced hotel renovations that guest B1 didn't. However, the within-platform identification strategy does not take into account the difference between reviews B1 and B2 to estimate the ATT.

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 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 an online travel agent that provides services like 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 some of our analyses we use Expedia ratings as a control group, we also create a subset of data that excludes 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.³

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 2 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

³We 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.

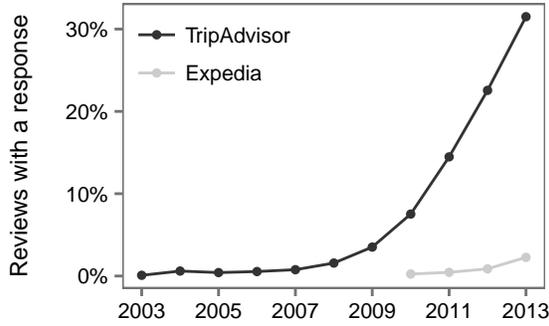


Figure 2: The cumulative percentage of reviews with a response by year.

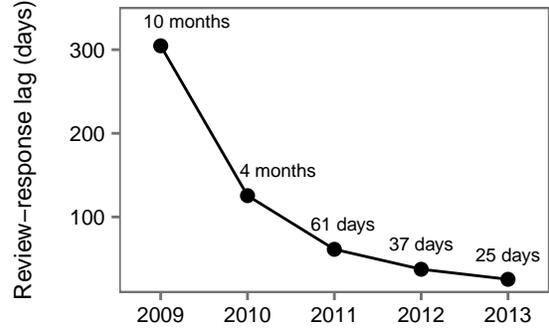


Figure 3: Average lag (in days) between a TripAdvisor review and its management response by review submission year.

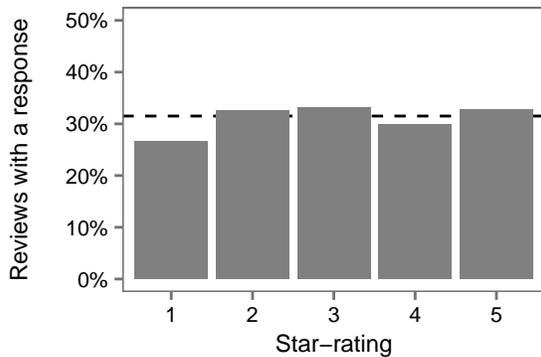


Figure 4: The fraction of TripAdvisor reviews that carry a response by star-rating. The overall average is 31.5% (dashed line.)

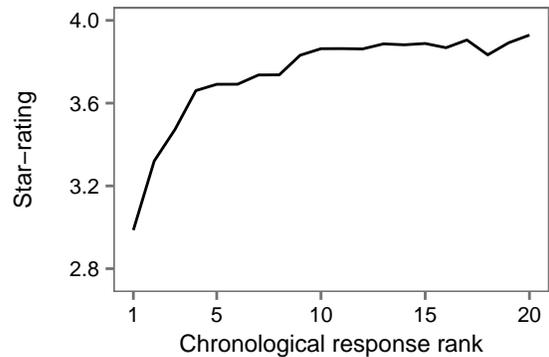


Figure 5: Average rating by chronological management response rank.

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 TripAdvisor. 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 3 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 4 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. While reviews with different ratings eventually receive responses at approximately the same rate, managers tend to respond to negative reviews first. We demonstrate this in Figure 5. The figure plots the average rating of reviews that received a management response, in chronological order. We see that while the first response goes to a review with an average rating of approximately 3 stars, the rating associated with the 20th response is nearly 4 stars. This pattern of responding causes a *transient* endogeneity problem: because managers tend to respond to negative reviews first, ratings following the adoption of management responses are likely to be higher than ratings submitted just before a manager’s first response regardless of any effect management responses may have on ratings.

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 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

$$\text{Stars}_{ijt} = \beta_1 \text{After}_{ijt} + \beta_2 \text{TripAdvisor}_j + \gamma \text{Interval}_{ijt} \times \text{TripAdvisor}_{ij} + X_{ijt} \gamma + \alpha_j + \tau_t + \epsilon_{ijt}, \quad (3)$$

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 γ , the vector of interaction coefficients associated with each interval, is the difference-in-differences estimate of interest. 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. Finally, our model includes calendar-month fixed-effects τ_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 α_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.⁴

We estimate the model in Equation 3 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 6. The figure plots the estimated values of the interval coefficients γ , 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 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

⁴The results of a pooled regression are not meaningfully different.

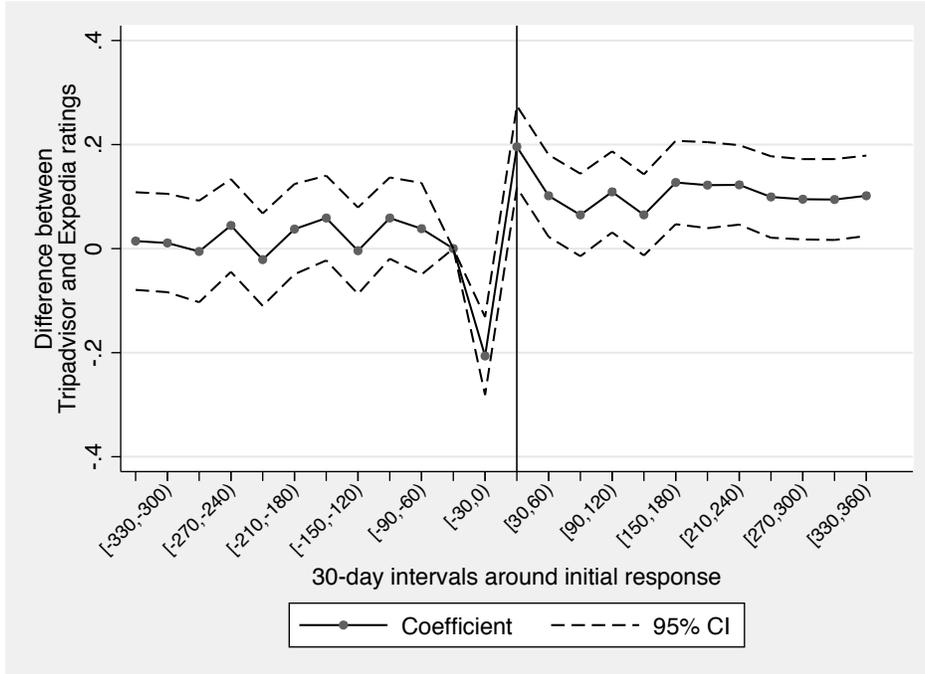


Figure 6: The evolution of treatment effects, *i.e.*, differences in hotel ratings between Expedia and TripAdvisor, as a function of a hotel’s decision to begin responding to reviews. The solid line plots the γ -coefficient estimates from Equation 3, and the dashed lines their respective 95% confidence intervals.

hypothesis of joint equality among pre-treatment intervals excluding $[-30, 0)$. Second, the figure reveals a negative outlier at $[-30, 0)$, which is caused by the fact that managers tend to respond to negative reviews first. While, on average, the adoption of management responses is preceded by a substantive negative shock to their TripAdvisor ratings, we do not know whether this association is causal. 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. Ashenfelter’s dip can be a sign of *transient* or *persistent* endogeneity.

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 *transient* endogeneity caused by Ashenfelter’s dip by computing long-run differences, where we 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:

$$\text{Stars}_{ijt} = \beta_1 \text{After}_{ijt} + \beta_2 \text{TripAdvisor}_{ij} + \delta \text{After}_{ijt} \times \text{TripAdvisor}_{ij} + X_{ijt} \gamma + \alpha_j + \tau_t + \epsilon_{ijt}, \quad (4)$$

where the variables are as in Equation 3, except that we replace the variable Interval_{ijt} for the variable After_{ijt} . Again, the matched-hotel fixed effects α_j ensure that our identification relies only on within hotel variation, *i.e.*, comparing the ratings of any given hotel on TripAdvisor with the ratings of the *same* hotel on Expedia. The primary coefficient of interest is δ , which measures the causal impact of management responses on hotel ratings.

We first estimate Equation 4 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 4. The estimated coefficient for the interaction term $\text{After}_{ijt} \times \text{TripAdvisor}_{ij}$ is 0.15 stars, and it is 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.⁵ We present these results in the second column of Table 4. As expected, our adjusted estimate for δ 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 After_{ijt} , which measures changes in the ratings of Expedia reviewers over the same time period is also of interest as it can be seen as treatment effect on the non-treated. 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 change whatsoever. We perform additional robustness checks against this type of concern in section 4.2.

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. In the next section, we perform robustness checks to verify that our results hold when accounting for various forms of endogeneity that cross-platform DD cannot address.

⁵Sensitivity tests excluding longer periods did not yield meaningfully different results.

4.2 Robustness checks for cross-platform DD

Differences in cross-platform traveler demographics and TripAdvisor-specific improvements. A key implication of the assumption underlying cross-platform DD identification is that TripAdvisor and Expedia users do not differentially value certain hotel improvements that happen to coincide with the adoption of management responses. If this assumption fails, cross-platform DD will lead to upwards biased estimates. To exemplify this concern, suppose that the dominant demographic on TripAdvisor is business travelers while there are few or no Expedia users who belong to this travel segment. Then, a hotel manager monitoring TripAdvisor reviews might simultaneously react in two ways. First, the manager might ensure that the concerns raised in the reviews of business travelers are addressed (*e.g.*, by making improvements to the hotel’s business center.) Second, the manager may respond to the TripAdvisor reviews that raised these concerns. Under these circumstances, the manager’s action could result in a TripAdvisor-specific increase ratings thereby inducing bias in our estimation.

How likely is this type of bias in our setting? Recall that previously we found that Expedia ratings do not change at all following the adoption of management responses on TripAdvisor (the coefficient for After_{ijt} is statistically indistinguishable from zero.) Therefore, if the effect we measure is due to unobserved hotel improvements, then Expedia users do not value these improvements at all. Even though it is plausible that Expedia users have different tastes than TripAdvisor users, and, indeed, that they value TripAdvisor-specific improvements less than TripAdvisor users, it is less likely that Expedia users’ tastes are so widely different that they do not value TripAdvisor-specific improvements at all. Nevertheless, we cannot rule out that Expedia users have zero value for TripAdvisor-specific improvements *and* hotels target their improvements at traveler segments that are overrepresented by a wide margin on TripAdvisor *and* that these TripAdvisor-specific improvements coincide with the adoption of management responses. In this section, we perform additional robustness checks to guard against this type of concern.

Our robustness checks rely on the fact that both TripAdvisor and Expedia ask reviewers about the purpose of their trip at review submission time. This information is voluntarily provided by reviewers, and therefore not all reviews carry such a designation. Moreover, in our sample, Expedia appears to have started collecting this information in 2010 whereas TripAdvisor started collecting this information as early as 2003. Nevertheless, the number of reviews carrying this label is substantial: considering post-2009 reviews, 48% of Expedia reviews and 89% of TripAdvisor reviews are associated with a particular traveler segment. The four most popular traveler segments, on both platforms, are “business”, “couples”, “families”, and “friends”. Expedia allows users to select among other less popular choices (such as “golfing trip” and “students”) that do not exist as options on TripAdvisor. We focus our analysis on the four segments that exist on both platforms, and which comprise the majority of labeled reviews. We then repeat our cross-platform DD estimation by traveler segment. The motivation for this robustness check is that by separately analyzing each traveler segment we lower the probability of bias arising from cross-platform reviewer heterogeneity. We present these results in Table 6. We find that our results are robust to conditioning on traveler

segment. Management responses have a positive (and, interestingly, similar in magnitude) impact on the ratings of the different traveler types. Taken together these by-segment regressions suggest that our results are unlikely to be due to TripAdvisor-specific improvements.

Changes in the review environment and reviewer selection A different type of concern with our results is that we have not accounted for changes in the review environment other than the adoption of management responses.⁶ A number of papers, including Godes and Silva (2012) and Moe and Schweidel (2012), discuss the role of the review environment consumers encounter on both the decision to leave a review as well as the review’s valence. If the timing of the adoption of management responses happens to coincide with changes in the review environment that result in increased ratings our estimates will be biased. In fact, as we have seen in Figure 6 hotels do adopt management responses following an unusually large negative shocks in their ratings, *i.e.*, a change in their review environment. Given the dynamic nature of changes in the review environment, the Ashenfelter’s dip correction we have used so far may not fully correct for this type of bias. For instance, consider the following hypothetical scenario. After a hotel receives a string a bad reviews, two things happen: a) the hotel starts responding, b) hotel guests who had a positive experience start inflating their ratings to compensate for what they perceive as inaccurately low prior ratings. In this case, it would be these “activist” reviewers causing the increase in ratings, and not the management responses.⁷ To test the robustness of our results to changes in review environment dynamics we include two salient characteristics of the review environment as controls in our cross-platform DD specification: for each review we compute (the log of) the number of TripAdvisor reviews preceding it and the average rating of these prior reviews.

We report these results in the third column of Table 4. The impact of management responses on ratings remains robust to the inclusion of review environment controls. However, some care is needed in interpreting the estimated coefficient for the treatment effect ($\text{After}_{ijt} \times \text{TripAdvisor}_{ij}$). While in some cases (like the one described in the previous paragraph) the inclusion of review environment controls will correct for unobserved bias, in other cases, including review environment controls could, in fact, introduce bias rather than correcting for it. Specifically, the ATT will be downward biased if the average rating of prior reviews positively affects future ratings. Prior empirical studies (*e.g.* Li and Hitt (2008)) find a positive association between average rating and subsequent reviews. This association can cause a feedback loop: a hotel manager responds to a review; in turn, this results in a subsequent positive review, which increases the hotel’s average rating; finally, the increased average rating itself raises the incidence of positive reviews. In this case, the average rating of prior reviews mediates the relationship between management responses and ratings. More generally this type of bias arises when management responses *cause* changes the review environment which then *cause* increases in ratings. However, even in such cases, there is a useful way to interpret the difference in the coefficients for the ATT in the presence and absence of

⁶We thank an anonymous reviewer for this suggestion.

⁷A priori, while this behavior is plausible, we think it is unlikely to persist over long periods. Presumably, once the “correction” happens reviewers will stop inflating their ratings.

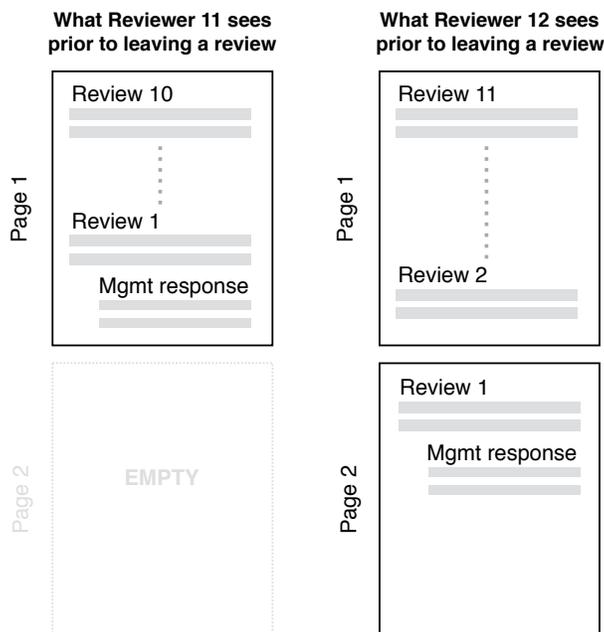


Figure 7: Identifying the impact of management responses by exploiting variation in the likelihood of reading a management response. Reviewer 11 is more likely to read the management response to Review 1 than Reviewer 12 is. By the time Reviewer 12 arrives to leave a review, the management response is displayed on page 2 and is thus less likely to be read.

the review environment controls (columns 2 and 3 of Table 4): their difference captures the indirect effect of management responses on ratings through their positive impact on a hotel's average rating.

Management response visibility as a treatment indicator Our analyses so far used management response adoption as a treatment indicator. Under this treatment scheme, all TripAdvisor reviews left after a hotel's first management response were part of the treatment group, while TripAdvisor reviews left prior to a hotel's first response were part of the control group. Then, we estimated an ATT by taking the difference in ratings between the treatment and control groups. If hotels took other unobserved actions that specifically affected their TripAdvisor ratings at the same time they started responding then this estimate could be biased. Consider for instance the case of TripAdvisor-specific hotel improvements: if hotels make improvements that are specifically appealing to TripAdvisor users at the same time they start responding, an ATT estimated as above will reflect both the impact of both management responses as well as the impact of these improvements. Here, we explicitly guard against this endogeneity concern, by identifying a control group of TripAdvisor users who were unlikely to be affected by management responses even though they reviewed hotels after they had started responding (and were thus affected by TripAdvisor-specific improvements or other unobserved hotel actions coinciding with the adoption of management responses.)

While we cannot precisely know which reviewers were exposed to management responses, we

can exploit the fact that TripAdvisor orders reviews by date and displays ten reviews per page to construct a proxy. As an example, which we illustrate in Figure 7, consider a hotel that has ten reviews, and that has only responded to the first review it received. Then, consider what the hotel’s next two reviewers, whom we label “Reviewer 11” and “Reviewer 12”, see. When Reviewer 11 arrives to leave a review (as shown in the left column of Fig. 7) the management response is still visible on the hotel’s first page of reviews. After Reviewer 11 leaves a review (as shown in the right column of Fig. 7), the review carrying the management response will be relegated to the hotel’s second page of reviews. Therefore, Reviewer 12 will be less likely to read the response than Reviewer 11. Because the effect of management responses should be larger for reviewers who are more likely to have read them, we can use reviewers like “Reviewer 12” as a control group.

Concretely, since reviewers are more likely to read the first page of reviews than they are to click through, and also read the second page of reviews, we construct the variable $\text{Pct. page 1 responded}_{ijt}$, which measures the fraction of the 10 most recent reviews (*i.e.*, the reviews on page 1) prior to review i that carried a response. We then interact this proxy variable with $\text{After}_{ijt} \times \text{TripAdvisor}_{ij}$ and re-estimate our model. We report these results in the second column of Table 7. We find a positive and significant interaction effect for $\text{Pct. page 1 responded}_{ijt}$. This suggests that reviewers who are more likely to read a management responses are more likely to be affected by them. Following the same logic, we construct the variable $\text{Pct. page 2 responded}_{ijt}$, which denotes the fraction of reviews on page 2 that carried a management responses at time review i was posted. We re-estimate the cross-platform DD model including interactions for both the page 1 and page 2 proxies. We report these results in the third column of Table 7. The estimate of the page 2 proxy is smaller and not statistically significant, coinciding with our intuition that users are less likely to be affected by management responses on the second page of a hotel’s reviews.

Finally, to reinforce the point that identification using management response visibility as a treatment indicator is not vulnerable to endogenous changes in ratings happening at the time hotels start responding, we estimate the same two specifications as in the previous paragraph using only reviews submitted following each hotel’s first response. The intuition behind this analysis is that if a hotel starts responding when it renovates then all subsequent reviewers experience these renovations. Therefore, while the difference between a rating submitted prior to a hotel’s first response and a rating submitted after a hotel’s response could be driven by unobserved TripAdvisor-specific improvements, it’s harder to argue the same for the difference between two ratings that are both submitted after a hotel begins responding. The results of these analyses, which we display in Table 8, are similar to our estimates using the entire dataset of reviews.

These robustness checks suggest that the effect we measure is due to management responses. Specifically, our results indicate that the impact of responding is higher in situations where management responses are more likely to have been read. By contrast, in situations where management responses are not displayed prominently (*e.g.*, on the second page of a hotel’s TripAdvisor reviews), their impact is smaller. Further, these results are unlikely to be explained by hotel renovations. While renovations are likely to drive increased ratings, we have less reason to believe that renova-

tions will differentially impact hotel guests depending on their likelihood of reading an management response *after* their stay. One limitation of the analyses in the section is that our response visibility proxy is almost certainly measured with error: some reviewers will not notice management responses on the first page of a hotel’s reviews, while other reviewers will notice management responses buried in a hotel’s last page of reviews. Such measurement error will attenuate the ATT we estimate.

Management responses and review fraud An identification concern arises if hotels that adopt management responses simultaneously adopt other reputation management strategies such as posting fake reviews. In this case, we may mistake increases in ratings due to review fraud for increases in ratings due to management responses, resulting in a positive bias in the ATT we estimate. Interestingly, the sign of such bias can also be negative. If hotels choose to stop posting fraudulent reviews when the option of directly responding to consumers becomes available to them, the ATT we estimate will be biased downwards. Therefore, while this type of bias is a concern, its direction will depend on whether management responses and review fraud are substitutes or complements. Whether management responses encourage or discourage review fraud activity is an interesting open question with implications for the design of review platforms. The cross-platform DD strategy is especially susceptible to review fraud biases because posting fake reviews is easier on TripAdvisor than it is on Expedia: while any traveler can leave a review on TripAdvisor, Expedia requires that users have paid-and-stayed.⁸

We perform two robustness to mitigate concerns arising from review fraud. Both checks rely on the fact that some firms have higher incentives to commit review fraud than others. If firms predisposed to review fraud are the ones that benefit from management responses, we might worry that review fraud is biasing our results.

For our first robustness check, we leverage the fact that review fraud incentives vary by hotel organizational form. Specifically, prior work (Mayzlin et al., 2014; Luca and Zervas, 2015) has shown that chain-affiliated firms are less likely to post fake reviews than independent firms. This difference in review fraud incentives arises for two reasons. First, because chain hotels benefit less from consumer reviews (Luca, 2011) they have weaker incentives to commit review fraud in the first place. Second, if a chain hotel is caught committing review fraud, there can be negative spillover effects on the reputation of the brand it is affiliated with. For this reason, as Mayzlin et al. (2014) point out, some chains have adopted social media policies that prohibit anyone other than their guests (*e.g.*, the chain’s employees) from posting reviews. Based on this observation, we repeat our analysis separately for independent and chain-affiliated hotels. We report these results in Table 9. Looking at chain hotels, which are unlikely to commit review fraud, we find that the impact of management responses on their ratings is positive, significant, and of similar magnitude to our

⁸Even though TripAdvisor allows anyone to post a review, it tries to ensure the integrity of the content that it publishes. For more, see http://www.tripadvisor.com/vpages/review_mod_fraud_detect.html. Therefore, not every fake review that is submitted to TripAdvisor will end up being published. Similarly, even though Expedia requires that consumers paid-and-stayed, review fraud is still possible: a hotel can create a fake reservation to allow it to post a fake review.

previous estimates (0.11, $p < 0.001$). This result suggests that the ATT we estimate is unlikely to be inflated due to review fraud. Intriguingly, we estimate a larger ATT (0.19) for non-chains. While it is tempting to interpret this result as evidence of independent hotel review fraud coinciding with the adoption of management responses, it could also be the case that management responses have a stronger impact on the reputation of independent hotels than the reputation of chains.

Our second robustness check relies on evidence from the literature suggesting that hotels with fewer reviews are more likely to commit review fraud to enhance their reputations (Luca and Zervas, 2015). At the same time, there is less reason for hotels with fewer reviews should benefit more from management responses. Therefore, if hotels with fewer reviews see greater increases in their ratings after they start responding, we might worry about confounding arising from review fraud.⁹ To test if the benefits from responding vary by pre-treatment review volume, we augment Equation 1 with an interaction term between treatment and the number of pre-treatment reviews for each hotel (*i.e.*, the number of reviews the hotel had just prior to its first response.) We report these results in the third column of Table 9. The interaction term is statistically indistinguishable from zero, suggesting that the impact of management responses is independent from the number of reviews a hotel had when it decided to start responding.¹⁰ This robustness checks provides additional evidence that benefits from responding do not vary by a hotel’s incentives to commit review fraud.

Difference in differences in differences As a final robustness check, we replicate our results using DDD, 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 controls 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 platform trends by differencing out cross-platform changes in the 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:

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

The following model implements our DDD estimator:

$$\begin{aligned} \text{Stars}_{ijt} = & \beta_1 \text{Responding}_j + \beta_2 \text{TripAdvisor}_{ij} + \beta_3 \text{Responding}_j \times \text{TripAdvisor}_{ij} & (5) \\ & + \beta_3 \text{Responding}_j \times \tau_t + \beta_3 \text{TripAdvisor}_{ij} \times \tau_t \\ & + \delta \text{After}_{ijt} \times \text{Responding}_j \times \text{TripAdvisor}_{ij} + X_{ijt} \gamma + \alpha_j + \tau_t + \epsilon_{ijt} \end{aligned}$$

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 δ . Because we can match TripAd-

⁹We thank K. Sudhir for suggesting this robustness check.

¹⁰Interacting with the log of pre-treatment responses also yields a zero coefficient.

visor to Expedia ratings, we use matched-pair fixed effects α_j , which subsume the coefficient for Responding_{*j*}. We report our results, first without and then with Ashenfelter’s dip correction, in Table 5. The DDD estimate (0.08 stars, $p < 0.01$) 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, supports our results so far.

Sensitivity analysis using Rosenbaum bounds Our cross-platform DD and DDD identification strategies use a 1-1 matched sample of treated and untreated units to identify the impact of management responses on hotel ratings. While matching the reviews of the same hotel across different platforms ensures compatibility in terms of observables, it does not mitigate the problem of selection on (time-varying) unobservables. Thus far, we dealt with selection on unobservables by performing case-specific robustness checks against hidden biases such as TripAdvisor-specific improvements and review fraud. Now, we assess the overall sensitivity of our estimates to any kind of hidden bias using Rosenbaum bounds (Rosenbaum, 2002).

One benefit of using Rosenbaum bounds is that we can assess the sensitivity of our results to hidden bias without having to specify how such bias might arise in practice. Specifically, suppose that treatment assignment (conditional on observables) is biased such that the odds of treatment of a unit and its matched control differ by a multiplier Γ , where $\Gamma = 1$ corresponds to the case of random treatment assignment. It is helpful to conceptualize such bias as the result of an unobserved covariate that both affects selection into treatment by a factor Γ *and* that is highly predictive of the outcome we are measuring. Because of this double requirement on the unobservable, Rosenbaum bounds are considered worst-case analyses (DiPrete and Gangl, 2004). Using Rosenbaum’s methods we can compute an upper bound on the p -value associated with the treatment effect assuming selection on unobservables of magnitude Γ .

We compute Rosenbaum bounds at various levels of Γ to examine how biases of different size would affect the significance level of the ATT. Because in our setting treatment is assigned to clusters (hotel-platforms) rather than individuals, we adjust our bounds for clustered treatment assignment (Hansen et al., 2014). Not accounting for clustering would exaggerate our effective sample size in a manner similar to using non-clustered standard errors. Table 16 displays upper bounds (p_{max}) on the p -value associated with the ATT at different levels of the sensitivity parameter Γ . We find that the minimum value of Γ at which the treatment effect we estimate becomes statistically insignificant at the 5% level is just below 4.5. The literature typically interprets values of $\Gamma > 2$ as evidence for robustness to large biases.

4.3 Results for within-platform identification

Arguably, the key concern with cross-platform identification is that differencing does not completely eliminate bias arising from unobserved differences between TripAdvisor and Expedia that may be correlated both with the adoption of management responses and changes in hotel ratings. Here, we use the within-platform identification strategy described in Section 2.2 to estimate the impact

of management responses. We implement this identification strategy with the following model:

$$\begin{aligned} \text{Stars}_{ijt} &= \beta_1 \text{Responding}_j + \delta \text{After}_{ijt} \times \text{Responding}_j + \mathbf{X}_{ijt} \gamma \\ &+ \eta_j \times \text{Year-Month Stayed}_{ijt} + \tau_t + \epsilon_{ijt}, \end{aligned} \tag{6}$$

where the interactions $\eta_j \times \text{Year-Month Stayed}_{ijt}$ are hotel-year-month-of-stay fixed effects. The precision of these fixed effects is at the year-month level because TripAdvisor does not disclose exact dates of travel, likely to protect user privacy. In total, our model contains over 110,000 such fixed effects in addition to time fixed-effects and linear time trends by treatment status. (Perhaps surprisingly, some variation remains in our data after we introduce all of these controls.) The effect of management responses is identified by variation in the difference between the ratings of TripAdvisor reviewers who left a review prior to a hotel’s adoption of management responses and the ratings of TripAdvisor reviewers who stayed at the same hotel during the same year-month but left a review following a hotel’s adoption of management responses.

While this identification strategy mitigates the concern of unobserved hotel renovations, bias can arise if the elapsed time between staying at a hotel and reviewing it is correlated with the guest’s rating. To account for endogeneity arising from review timing we include as controls the time elapsed between a review and a stay, as well as the square of the same variable (to allow for non-linear effects.) We report these results in the first column of Table 10. In the second column, we also correct for Ashenfelter’s dip to account for the fact that hotels tend to start responding when they experience negative shocks to their ratings. We find a positive and significant effect for responding whose magnitude is similar to our results so far.

A concern with using a flexible polynomial trend to absorb correlation between how long guests wait to leave a review and how enjoyable their stay was, is that the relationship between the two variables may be more complex. To avoid parametric assumptions about the relationship between rating and elapsed time, we would like the elapsed time covariate to be balanced between the treatment and control groups, *i.e.*, we’d like to have: $P(\text{Treated}|\text{Elapsed time between staying and reviewing}) = P(\text{Treated})$. Using management response visibility as the treatment indicator achieves this goal. A Kolmogorov-Smirnov test fails to reject the null hypothesis that treated and control reviewers have different distributions of elapsed times between staying and reviewing. Table 11 reports our within-platform estimates using management response visibility as a treatment indicator. As before, we confirm that the impact of management responses is stronger for reviewers who are more likely to have read them.

4.4 Robustness to alternative functional forms

In our analysis so far, we have modeled an ordered discrete outcome (the 1 to 5-star rating associated with each review) using a continuous linear model. While this modeling choice is common in the reviews literature, it misrepresents the data generation process and can lead to bias. In this section, we repeat our analysis using a generalized ordered probit specification, which reflects our

data generating process more accurately. We begin by briefly describing the generalized ordered probit model – for a complete description see (Terza, 1985)). The model posits that the cumulative probabilities of the discrete outcomes (the star ratings) are given by:

$$Pr[\text{Stars}_{ijt} \leq s \mid x_{ijk}, z_{ijk}] = \Phi(\kappa_s + x'_{ijk}\beta_s + z'_{ijk}\gamma) \quad s = 1 \text{ to } 4, \quad (7)$$

where Φ is the cumulative normal distribution. Compared to the standard ordered probit, the generalized model allows some of its coefficients (the β_s) to vary by outcome. This generalization relaxes the parallel regressions assumption of the standard ordered probit model, and allows the effect of covariates to vary across outcomes. We begin by estimating the generalized ordered probit model on the TripAdvisor ratings of responding hotels. In the set of threshold-varying controls, we include an indicator After_{ijt} denoting the post-responses period. In addition, to flexibly control for unobserved time trends, we also include a set of year dummies and linear time trends (whose coefficients do not vary by outcome to avoid introducing too many parameters in the model.)

We estimate the model using MLE and compute standard errors clustered at the hotel level with a non-parametric bootstrap. We report our results in the first column of Table 12. While these estimates are not as easily interpretable as in the linear case, in general, a set of positive and significant coefficients (as we find here) suggest an increase in the probability of higher ratings. To arrive at more interpretable estimates we also compute average marginal probability effects (MPE's) as described in Boes and Winkelmann (2006). Omitting irrelevant subscripts for simplicity, marginal probability effects are defined as:

$$\text{MPE}_{sl}(x) = \partial Pr[\text{Stars} \leq s \mid x, z] / \partial \beta_s^{(l)} = \phi(\kappa_s + x'\beta_s)\beta_s^{(l)} - \phi(\kappa_{s-1} + x'\beta_s)\beta_{s-1}^{(l)} \quad (8)$$

where $\beta_s^{(l)}$ denotes l^{th} item of the vector β_s . Then, the average MPE's are defined as $E_x[\text{MPE}_{sl}(x)]$, and they should be interpreted as average probability changes given a marginal change in the covariate of interest. Average MPE's can be consistently estimated using the estimated model parameters in place of the true parameters. We report average MPE's and bootstrap standard errors (clustered at the hotel level) for After_{ijt} in the first column of Table 13. We find that the likelihood of receiving a 5-star review increases by approximately 7% following the adoption of management responses. Meanwhile, the probability of a 1-star rating decreases by nearly 2%. These results are in line with our previous DD estimates using a linear model.

In the spirit of DD, we also perform a falsification check. Specifically, we re-estimate the same generalized ordered probit model on the Expedia reviews of these same hotels that respond on TripAdvisor. Here, we set the variable After_{ijt} to 1 for all Expedia reviews following each hotel's first management response on TripAdvisor. We report these estimates and their associated average MPE's in the second column of Tables 12 and 13. As expected, we find no change in the Expedia ratings of responding hotels following their adoption of management responses on TripAdvisor.

5 Why do management responses affect hotel ratings?

In this section, we investigate the mechanism underlying our findings. We argue that management responses can improve hotel ratings because they increase the cost of leaving a negative review while making it more worthwhile to leave a positive one. Intuitively, the cost of negative reviews increases because when hotels respond, consumers feel that their reviews will be closely scrutinized. Therefore, consumers become less likely to submit low quality negative reviews. On the other hand, consumers considering leaving a positive review likely appreciate the hotel reading their review and responding to them. Therefore, hotel guests are more likely to submit a positive review when hotels take notice of their feedback.

To empirically support this argument, we analyze the impact of management responses on review volume, review length, and the types of reviewers a hotel attracts. Beyond helping us understand the mechanism underlying our findings, these analyses yield insights on managerially relevant variables other than star-ratings.

Our first finding is that the length of negative reviews tends to increase after hotels begin responding. To arrive at this result, we employ the same cross-platform DD strategy used in Section 4.1. Thus, we estimate Equation 1, but using the review length (measured in characters) as dependent variable. Negative reviews on TripAdvisor are, on average, longer than positive reviews. Therefore, we separately estimate the impact of management responses on review length for each star-rating and report these results in columns 2–6 of Table 15. Because the average TripAdvisor rating of responding hotels is 3.8 stars, we define negative reviews as those with 1, 2 or 3 stars, and positive reviews as those with 4 or 5 stars. We find that reviewers leave 1 and 2-star reviews that are approximately 10% longer after hotels begin responding. The impact on 3-star reviews is smaller, while the length of positive reviews remains unchanged. Thus, we find that hotel managers who consider responding to reviews face an interesting trade-off: by responding they can increase their average star rating at the cost of receiving longer, and therefore more detailed, negative reviews.

More interestingly, this finding can also help us explain *why* management responses increase hotel ratings. Hotel guests feel the need to leave longer and more detailed reviews when they believe that hotel managers will scrutinize their comments and publicly respond. For some guests, writing a longer and more detailed negative review will be worth the time and effort. Others, however, will not be motivated to expend this extra effort, and instead will opt for not leaving any review at all. In other words, management responses increase the cost of writing a negative review.

Second, we find that following a hotel’s decision to begin responding, total review volume increases. Since on average ratings also increase, these extra reviews are mostly positive. Again, we estimate the impact on review volume using the cross-platform DD strategy (Equation 1). Specifically, we estimate the percentage change in the number of reviews a hotel receives after it begins responding on TripAdvisor, relative to percentage increases on Expedia over the same period of time. To do so, we first aggregate our data at the hotel-month level. Then, our dependent variable is $\log \text{Review count}_{jt}$, *i.e.*, the logarithm of the number of reviews hotel j received in month t . As before we cluster errors at the hotel level. We report these results in the first column of Table 15.

We find that the number of reviews a hotel receives increases by 12% following its decision to begin responding.¹¹ Why does review volume increase? Plainly, we think that positive reviewers who might have otherwise not left a review, are more willing to provide feedback when the hotel has signaled that it is listening. We also point out that, all else equal, an increased number of reviews is a desirable outcome because it is often interpreted as sign of hotel popularity and, thus, quality.

Third, we argue that if there is an increased benefit of leaving positive reviews when hotels respond, then reviewers who are inherently more positive should review the hotel more often. We define an inherently positive reviewer as someone who tends to leave more positive reviews than the average TripAdvisor reviewer, whether a firm responds or not. To show that responding hotels attract more inherently positive reviewers, we begin with the observation that ratings can be decomposed into three components: a reviewer fixed effect θ_k that captures how positive a reviewer is on average; a hotel fixed effect η_j that captures the average quality of hotel j ; and, an idiosyncratic shock ϵ_{jk} .¹² Then, the rating of reviewer k for business j is given by

$$\text{Stars}_{jk} = \theta_k + \eta_j + \epsilon_{jk}. \quad (9)$$

We estimate the reviewer fixed effects θ_k based on a hold-out set of reviews that contains each reviewer’s entire TripAdvisor review history excluding reviews for responding hotels.

Then, to test the hypothesis that when hotels start responding they attract reviewers who are inherently more positive, we estimate the following model using the TripAdvisor reviews of both responding and non-responding hotels:

$$\text{Reviewer type}_{ijt} = \beta \text{After}_{ijt} + \eta_j + \tau_t + \epsilon_{jk}. \quad (10)$$

Here, the dependent variable $\text{Reviewer type}_{ijt}$ is the value of θ_k associated with the reviewer who wrote review i for hotel j (as estimated using Equation 9.) After_{ijt} is an indicator for reviews submitted after hotel j starts responding. The coefficient of interest, β , captures changes in reviewer positivity after hotels start responding. To further limit the influence of unobserved transient factors that could affect reviewer selection, we limit our estimation sample to one year before and after the treatment, since any two reviewers are more likely to be comparable in their unobserved characteristics if their reviews are closer in time. We present our results in Table 14. We find that reviewers who leave reviews after hotels start responding are, on average, 0.04 stars more positive than reviewers who left reviews prior to the adoption of management responses. A robustness check using 6-month bandwidth, shown in the second column of Table 14, yields similar results.

This finding further supports the idea that management responses directly affect reviewer selection: once hotels start responding they attract reviewers who are inherently more positive in their evaluations regardless of whether hotels respond, or not.

¹¹A fixed-effects Poisson model gives a similar estimate.

¹²Dai 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.

5.1 Management responses and retaliatory reviewing

We briefly highlight a theoretical connection between our results and the literature on retaliation in bilateral review platforms. A number of field and lab studies (Resnick and Zeckhauser, 2002; Dellarocas and Wood, 2008; Bolton et al., 2013) have shown that in settings where agents can sequentially rate each other negative ratings are underreported, because of a fear of retaliation. 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 a 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). More recently, Airbnb has faced similar issues (Fradkin et al., 2014; Zervas et al., 2015).

Here, we draw a parallel between management responses and bilateral reviewing: hotels can “retaliate” negative reviews by disputing a reviewer’s claims in a management response, which in turn may discourage future guests with a negative experience from leaving a review altogether. This behavior can shift reviewer selection towards reviewers with higher ratings, and, on average, improve the ratings of responding hotels. A limitation of using the retaliation theory to explain our findings is that, unlike bilateral review platform, TripAdvisor does not allow hotels to rate their guests, which would visibly harm the guests’ online reputation. Thus, the main risk a reviewer faces in leaving a negative TripAdvisor review is primarily psychological. While the direct economic consequences of an antagonistic management response are not clear, some existing research (Ockenfels et al., 2012) suggests that consumers place more value on their online reputation than economic incentives alone would predict. For instance, the threat of an antagonistic management response may incur social and emotional costs that can affect a reviewer’s decision to leave a negative review.

5.2 Other mechanisms to explain our findings

A change in reviewing costs is not the only potential explanation for our results. Here, we briefly discuss a second mechanism that could in principle explain our findings, but find limited evidence to back it up. 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 for this hypothesis in our data, which we present in detail in the appendix. However, the number of reviews by returning consumers is too small (1.3% of all TripAdvisor reviews) to adequately explain the increase in ratings of responding hotels. As the number of reviews by returning consumers grows, this will be a hypothesis worth revisiting.

6 Managerial implications & conclusion

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 hypothesize that management responses increase the cost of leaving a negative review, while decreasing the cost of leaving a positive one. We empirically support this hypothesis by showing that following the adoption of management responses negative reviews become longer (*i.e.*, costlier to produce), overall review volume increase, and hotels attract reviewers who are inherently more positive in their evaluations.

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). One downside of responding is that hotels are more likely to attract fewer but more detailed negative reviews from guests who are trying harder to substantiate their complaints knowing that hotels will scrutinize their feedback. This highlights an interesting trade-off for managers. Our own experience as consumers, often focusing on reading negative reviews first, suggests that the risks in longer negative reviews may in some instances outweigh the benefits of increased ratings. Quantifying these trade-offs is an interesting area for future research.

A limitation to the conclusion that management responses can help firms improve their ratings is that our work does not estimate the impact of using management responses for a *randomly* chosen hotel, *i.e.*, we estimate at ATT instead of an ATE. Despite this limitation, we see two significant implications that we can draw from the ATT. First, our work informs hotels that are currently responding to reviews, about the effects of management responses on their reputations, an effect they may not have been aware of. Second, even though the treatment effect could be significantly different for hotels that do not currently respond, we speculate that this is unlikely to be the case. Our analysis indicates that the primary driver of improved reputation is a change in reviewer behavior rather than any particular hotel characteristic. Further, in many instances, responding and non-responding hotels are highly similar: we can match approximately 25% of non-responding chains to a responding chain with the same affiliation in the same city. For instance, while Americas Best Value Inn at 3243 Merrifield Avenue, Dallas currently responds the Americas Best Value Inn at 4154 Preferred Place, Dallas doesn't. This is an example where we might expect the impact of management responses to be similar for the two hotels. Therefore, even though our results should not be taken as a definite prescription for improving a firm's online reputation, we think that management responses are promising reputation management strategy even for hotels that not currently responding.

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. Further, as we have shown management responses, encourage review creation. Therefore, management responses can help review platforms grow the size of their

review collections, which is a metric review platforms commonly use to evaluate their success. On the other hand, our work shows that management responses have the undesired consequence of negative review underreporting, which positively biases the ratings of responding hotels. This is a downside for review platforms striving to maintain unbiased ratings, and for consumers who might be misled.

Our results also have implications for review platforms that do not allow responding, or for platforms like Expedia on which hotels tend not to respond. As we have shown, management responses lead to more reviews. But, where do these reviews come from? One possibility is that reviewers who would not have otherwise left a review, now choose to leave one. A more intriguing hypothesis is that management responses result in cross-platform substitution: reviewers migrate from platforms that do not allow management responses to platforms that allow management responses because their reviews are more likely to have an impact in the latter. Fully understanding the mechanism that drives review volume increases following the adoption of management responses is an interesting open question.

Taken together, our results 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.

Our results can be extended in various ways. For instance, managers who are considering responding to consumer reviews face 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 techniques to estimate heterogeneous treatment effects 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. A randomized field experiment to measure differences between the ATT and the ATE would be another interesting extension of our work. Such an experiment would help us understand if management responses work better for some firms than they do for others.

References

- O. C. Ashenfelter and D. Card. Using the longitudinal structure of earnings to estimate the effect of training programs, 1984.
- M. Bertrand, E. Duflo, and S. Mullainathan. How much should we trust differences-in-differences estimates? *Quarterly Journal of Economics*, 119(1), 2004.
- S. Boes and R. Winkelmann. Ordered response models. In *Modern Econometric Analysis*, pages 167–181. Springer, 2006.

- G. Bolton, B. Greiner, and A. Ockenfels. Engineering trust: reciprocity in the production of reputation information. *Management Science*, 59(2):265–285, 2013.
- J. A. Chevalier and D. Mayzlin. The effect of word of mouth on sales: Online book reviews. *Journal of marketing research*, 43(3):345–354, 2006.
- W. Dai, G. Z. Jin, J. Lee, and M. Luca. Optimal aggregation of consumer ratings: an application to yelp.com. Technical report, National Bureau of Economic Research, 2012.
- C. Dellarocas and C. A. Wood. The sound of silence in online feedback: Estimating trading risks in the presence of reporting bias. *Management Science*, 54(3):460–476, 2008.
- T. A. DiPrete and M. Gangl. Assessing bias in the estimation of causal effects: Rosenbaum bounds on matching estimators and instrumental variables estimation with imperfect instruments. *Sociological methodology*, 34(1):271–310, 2004.
- S. G. Donald and K. Lang. Inference with difference-in-differences and other panel data. *The review of Economics and Statistics*, 89(2):221–233, 2007.
- A. Fradkin, E. Grewal, D. Holtz, and M. Pearson. Reporting Bias and Reciprocity in Online Reviews: Evidence From Field Experiments on Airbnb. 2014. Working paper. Cited with permission. Available at http://andreyfradkin.com/assets/long_paper.pdf.
- T. Friedrich and M. L. Ordning. Effects of job-training programs on earnings. 2013.
- D. Godes and J. C. Silva. Sequential and temporal dynamics of online opinion. *Marketing Science*, 31(3):448–473, 2012.
- B. B. Hansen, P. R. Rosenbaum, and D. S. Small. Clustered treatment assignments and sensitivity to unmeasured biases in observational studies. *Journal of the American Statistical Association*, 109(505):133–144, 2014.
- J. J. Heckman and J. A. Smith. The pre-programme earnings dip and the determinants of participation in a social programme. implications for simple programme evaluation strategies. *The Economic Journal*, 109(457):313–348, 1999.
- X. Hui, M. Saeedi, Z. Shen, and N. Sundaresan. From lemon markets to managed markets: the evolution of ebays reputation system. 2014.
- C. Jepsen, K. Troske, and P. Coomes. The labor-market returns to community college degrees, diplomas, and certificates. *Journal of Labor Economics*, 32(1):95–121, 2014.
- X. Li and L. M. Hitt. Self-selection and information role of online product reviews. *Information Systems Research*, 19(4):456–474, 2008.
- X. Li, C. Gan, and B. Hu. The welfare impact of microcredit on rural households in china. *The Journal of Socio-Economics*, 40(4):404–411, 2011.

- M. Luca. Reviews, reputation, and revenue: The case of yelp. com. Technical report, Harvard Business School, 2011.
- M. Luca and G. Zervas. Fake it till you make it: Reputation, competition, and yelp review fraud. *Management Science*, *Forthcoming*, 2015.
- D. Mayzlin, Y. Dover, and J. Chevalier. Promotional reviews: An empirical investigation of online review manipulation. *American Economic Review*, 104(8):2421–55, 2014.
- M. A. McCollough, L. L. Berry, and M. S. Yadav. An empirical investigation of customer satisfaction after service failure and recovery. *Journal of service research*, 3(2):121–137, 2000.
- W. W. Moe and D. A. Schweidel. Online product opinions: Incidence, evaluation, and evolution. *Marketing Science*, 31(3):372–386, 2012.
- A. Ockenfels, P. Resnick, G. Bolton, and R. Croson. Negotiating reputations, 2012.
- P. Resnick and R. Zeckhauser. Trust among strangers in internet transactions: Empirical analysis of ebay’s reputation system. *Advances in applied microeconomics*, 11:127–157, 2002.
- P. R. Rosenbaum. *Observational studies*. Springer, 2002.
- A. K. Smith, R. N. Bolton, and J. Wagner. A model of customer satisfaction with service encounters involving failure and recovery. *Journal of marketing research*, pages 356–372, 1999.
- S. S. Tax, S. W. Brown, and M. Chandrashekar. Customer evaluations of service complaint experiences: implications for relationship marketing. *The Journal of Marketing*, pages 60–76, 1998.
- J. V. Terza. Ordinal probit: a generalization. *Communications in Statistics-Theory and Methods*, 14(1):1–11, 1985.
- G. Zervas, D. Proserpio, and J. Byers. A first look at online reputation on airbnb, where every stay is above average. *Where Every Stay is Above Average (January 23, 2015)*, 2015.

Table 1: Dataset description.

	TripAdvisor	Expedia
<i>All hotels</i>	5,356	3,845
Reviewed	4,603	3,356
Responding	2,590	587
Reviews	314,776	519,962
Responses	99,178	11,781
<i>Matched hotels</i>	3,681	3,681
Reviewed	3,511	3,265
Responding	2,387	568
Reviews	296,138	515,227
Responses	97,684	11,779
<i>Matched hotels reviewed on both platforms</i>	3,264	3,264
Responding	2,303	567
Reviews	291,119	515,223
Responses	96,665	11,776
<i>Cross-platform DD hotels</i> ¹³	1,762	1,762
Reviews	166,152	263,804
Responses	55,684	–
<i>Cross-platform DDD hotels</i> ¹⁴	2,697	2,697
Reviews	203,068	348,983
Responses	55,684	–

¹³Matched responding hotels that are reviewed on both platforms, excluding hotels that respond on Expedia

¹⁴Matched hotels that are reviewed on both platforms, excluding hotels that respond on Expedia

Table 2: Hotel summary statistics. A matched hotel is one which exists on both TripAdvisor and Expedia.

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

Table 3: Overview of the main identification strategies and robustness checks we perform. The “Pre” and “Post” datasets respectively indicate reviews submitted prior to and following each hotel’s first management response. All effect sizes reported below are corrected for Ashenfelter’s dip and are statistically significant at least at the 5% level.

Strategy	Treatment	Data used				Effect
		TripAdvisor	Expedia	Pre	Post	
Cross-platform DD	Response adoption	✓	✓	✓	✓	0.12
<i>By traveller segment</i>		✓	✓	✓	✓	
Business						0.09
Couples						0.18
Families						0.10
Friends						0.11
<i>By hotel affiliation</i>		✓	✓	✓	✓	
Non-chain						0.19
Chain						0.11
Cross-platform DD	Response visibility	✓	✓	✓	✓	0.08
		✓	✓		✓	0.07
Cross-platform DDD	Response adoption	✓	✓	✓	✓	0.08
Within-platform DD	Response adoption	✓		✓	✓	0.12
Within-platform DD	Response visibility	✓		✓	✓	0.07

Table 4: Cross-platform DD.

	(1)	(2)	(3)
After \times TripAdvisor	0.149*** (7.21)	0.123*** (5.49)	0.097*** (5.20)
TripAdvisor	-1.006*** (-20.38)	-1.027*** (-20.21)	-0.803*** (-18.31)
After	-0.005 (-0.45)	-0.012 (-0.91)	-0.003 (-0.24)
Avg. Rating			0.288*** (26.53)
Log Review Count			-0.003 (-0.69)
Ashenfelter's dip correction	No	Yes	Yes
N	429956	415361	411993
R ² within	0.020	0.020	0.024

Note: The dependent variable is rating i 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 linear time trends.

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

Table 5: Cross-platform DDD.

	(1)	(2)
After \times Responding \times Tripadvisor	0.113*** (6.59)	0.081*** (4.31)
Tripadvisor	0.923 (0.96)	0.896 (0.93)
Responding $\times \tau_t$	Yes	Yes
TripAdvisor $\times \tau_t$	Yes	Yes
Ashenfelter's dip correction	No	Yes
N	552051	537456
R ² within	0.021	0.021

Note: The dependent variable is rating i 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 linear time trends.

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

Table 6: Cross-platform DD by traveler segment.

	(1) Business	(2) Couples	(3) Families	(4) Friends
After \times Tripadvisor	0.093** (2.43)	0.176*** (4.54)	0.104*** (2.71)	0.111* (1.74)
Tripadvisor	-0.846*** (-4.19)	-0.520*** (-3.56)	-1.223*** (-7.99)	-0.695 (-1.28)
After	0.005 (0.15)	-0.066* (-1.88)	-0.025 (-0.78)	0.019 (0.30)
Ashenfelter's dip correction	Yes	Yes	Yes	Yes
N	59886	41126	62282	14787
R ² within	0.0068	0.016	0.017	0.021

Note: The dependent variable is rating i 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 linear time trends.

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

Table 7: Cross-platform DD using management response visibility as a treatment indicator.

	(1)	(2)
After \times Tripadvisor	0.101*** (4.00)	0.100*** (3.92)
After \times Tripadvisor \times Pct. page 1 responded	0.084*** (4.08)	0.062*** (2.76)
After \times Tripadvisor \times Pct. page 2 responded		0.013 (1.34)
Tripadvisor	-1.014*** (-19.95)	-1.012*** (-19.91)
After	-0.013 (-0.99)	-0.012 (-0.97)
Ashenfelter's dip correction	Yes	Yes
N	415361	415361
R ² within	0.020	0.020

Note: The dependent variable is rating i 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 linear time trends.

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

Table 8: Cross-platform DD using management response visibility as a treatment indicator only with reviews submitted after each hotel’s first management response.

	(1)	(2)
Tripadvisor	-0.789*** (-7.71)	-0.786*** (-7.69)
Tripadvisor \times Pct. page 1 responded	0.071*** (3.66)	0.056** (2.52)
Tripadvisor \times Pct. page 2 responded		0.009 (0.95)
Ashenfelter’s dip correction	Yes	Yes
N	274200	274200
R ² within	0.0097	0.0097

Note: The dependent variable is rating i 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 linear time trends.

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

Table 9: Cross-platform DD robustness checks for fake reviews: ATT by hotel affiliation and pre-treatment review volume.

	(1) Non-chain	(2) Chain	(3) By review volume
After \times TripAdvisor	0.195*** (2.65)	0.104*** (5.29)	0.126*** (5.33)
TripAdvisor	-1.016*** (-7.61)	-1.043*** (-21.10)	-1.026*** (-19.70)
After	-0.032 (-0.74)	-0.009 (-0.68)	-0.011 (-0.86)
After \times TripAdvisor \times Pre-treatment num. reviews			-0.000 (-0.69)
Ashenfelter’s dip correction	Yes	Yes	Yes
N	65902	349459	404231
R ² within	0.020	0.020	0.020

Note: The dependent variable is rating i 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 linear time trends.

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

Table 10: Within-plaform identification: Comparing the TripAdvisor ratings of travelers who stayed at the same hotel the same month.

	(1)	(2)
After	0.276*** (8.02)	0.121** (1.97)
Time between review & stay	0.037*** (4.56)	0.037*** (4.54)
Time between review & stay ²	-0.001*** (-4.15)	-0.001*** (-4.26)
Ashenfelter's dip correction	No	Yes
N	308261	299295
R ² within	0.0029	0.0025

Note: The dependent variable is rating i of hotel j at time t . Cluster-robust t-statistics (at the individual hotel-month level) are shown in parentheses. All specifications include hotel-month-of-stay fixed effects, time fixed effects and treatment-status specific linear time trends.

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

Table 11: Within-plaform identification using management response visibility as the treatment indicator.

	(1)	(2)
After	0.101 (1.64)	0.099 (1.61)
After \times Pct. page 1 responded	0.067*** (3.89)	0.056*** (2.63)
After \times Pct. page 2 responded		0.010 (0.89)
Time between review & stay	0.038*** (4.55)	0.038*** (4.56)
Time between review & stay ²	-0.001*** (-4.27)	-0.001*** (-4.27)
Ashenfelter's dip correction	Yes	Yes
N	299295	299295
R ² within	0.0026	0.0026

Note: The dependent variable is rating i of hotel j at time t . Cluster-robust t-statistics (at the individual hotel-month level) are shown in parentheses. All specifications include hotel-month-of-stay fixed effects, time fixed effects and treatment-status specific linear time trends.

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

Table 12: Generalized ordered probit.

	(1) TripAdvisor	(2) Expedia
Threshold 1 2		
After	0.168*** (3.76)	0.041 (1.26)
Threshold 2 3		
After	0.141*** (3.46)	0.033 (0.99)
Threshold 3 4		
After	0.145*** (3.24)	0.022 (0.63)
Threshold 4 5		
After	0.165*** (3.57)	0.019 (0.69)
Ashenfelter's dip correction	Yes	Yes
N	159772	255589

Note: The dependent variable is rating i of hotel j at time t . Bootstrap standard errors shown in parentheses. All specifications include year fixed effects and linear time trends. *Significance levels:* * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 13: Average marginal probability effects of generalized ordered probit.

	(1) TripAdvisor	(2) Expedia
1 star	-0.022*** (-4.04)	-0.004 (-1.27)
2 stars	-0.009** (-2.39)	-0.002 (-0.67)
3 stars	-0.016*** (-2.62)	-0.000 (-0.07)
4 stars	-0.018*** (-3.03)	-0.001 (-0.29)
5 stars	0.065*** (3.57)	0.008 (0.69)
Ashenfelter's dip correction	Yes	Yes
N	159772	255589

Note: Bootstrap standard errors shown in parentheses. All specifications include year fixed effects and linear time trends. *Significance levels:* * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 14: Change in reviewer types following a hotel’s decision to begin responding.

	(1) BW= ±12 months	(2) BW= ±6 months
After	0.040*** (4.47)	0.033*** (3.06)
N	59710	33284
R ² within	0.00061	0.00043

Note: The dependent variable is the reviewer type θ_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.

Table 15: The impact of management responses on reviewing activity and review length.

	(1)	(2)	(3)	(4)	(5)	(6)
	Num. reviews	Review length				
		1-star	2-stars	3-stars	4-stars	5-stars
After × Tripadvisor	0.12*** (4.57)	88.35*** (3.88)	93.24*** (3.83)	47.92*** (2.79)	19.60 (1.51)	8.28 (0.51)
Tripadvisor	−0.68*** (−14.78)	849.81*** (17.57)	1021.08*** (21.72)	981.80*** (24.64)	890.98*** (27.62)	717.68*** (17.64)
After	0.01 (0.70)	−0.50 (−0.03)	−10.84 (−0.87)	−13.63 (−1.62)	−6.13 (−1.18)	−10.84* (−1.73)
Ashenfelter’s dip correction	Yes	Yes	Yes	Yes	Yes	Yes
N	122350	22754	28427	51300	120319	192561
R ² within	0.24	0.16	0.18	0.19	0.21	0.21

Note: The dependent variable in column (1) is the log of the number of reviews of hotel j at time t . The dependent variable in columns (2-6) is the length of review i 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 linear time trends.

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

Table 16: Rosenbaum bounds for cross-platform DD.

Sensitivity parameter Γ	Maximum significance level p_{max}
1.0	0.000
1.5	0.000
2.0	0.001
2.5	0.004
3.0	0.010
3.5	0.020
4.0	0.034
4.5	0.051
5.0	0.070
5.5	0.091
6.0	0.114

A Management responses and service recovery

In this section, we show that reviewers who received a management response after leaving a negative review were 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

$$\text{Returning consumer}_{kj} = \beta \text{Received response}_{kj} + \eta_j + \epsilon_{kj}, \quad (11)$$

where $\text{Returning consumer}_{kj}$ is an indicator variable that is set to 1 if consumer k has left more than one review for hotel j , $\text{Received response}_{kj}$ is an indicator variable set to 1 for consumers who received a management response for their initial review of hotel j , and η_j is a hotel fixed effect. We present our results in the first column of Table 17. We find that consumers who received a management response are 9% more likely to provide a second review than 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. By limiting our analysis to 1- and 2-star reviews we also exclude many reviewers who return to the hotel, not because of the management response but because the failure they experienced was not that severe. Our results, in the second column of Table 17, 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.¹⁵ While we might expect any returning hotel guest to anticipate a better second experience, we isolate the impact of a 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

¹⁵Even though consumers can write more than two reviews for the same hotel, few consumers in our data provide three or more reviews for the same establishment. For simplicity, we focus on the first and second reviews of each consumer for a specific hotel. Specifically, 3,468 consumers return at least once to the same hotel, for a total of 8,151 reviews. Excluding their first reviews, these consumers produced 4,418 reviews (1.4% of total TripAdvisor reviews). 501 of these consumers return more than twice. Excluding their first and second reviews, these consumers wrote an additional 685 reviews (0.2% of total TripAdvisor reviews).

of every consumer who left at least two reviews, and we estimate the following regression model:

$$\begin{aligned} \text{Stars}_{ikj} = & \gamma_1 2^{nd} \text{ review}_{ikj} + \gamma_2 \text{Received response}_{kj} \\ & + \beta \text{Received response}_{kj} \times 2^{nd} \text{ review}_{ikj} + \eta_j + \epsilon_{ikj}. \end{aligned} \quad (12)$$

The dependent variable is the i^{th} rating of consumer k for hotel j , $\text{Received response}_{kj}$ is an indicator for consumers who received a response for their first review of hotel j , and $2^{nd} \text{ review}_{ikj}$ 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 β has the standard DD interpretation. Our results, shown in Table 18, indicate that returning consumers are more positive by almost 1 star, but those who receive a response increase their second ratings by more than half a star (0.57, $p < 0.05$), highlighting the effect of recovery efforts. Unfortunately, as indicated by the small sample size of this analysis ($N = 358$), the aggregate effect of such improved ratings on hotel reputation is insignificant. In fact, our main results remain practically unchanged when we exclude returning reviewers from our data. 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.

Table 17: Logistic regression of the probability of a second review by the same consumer as a function of receiving a management response.

	(1) All Reviews	(2) Only stars < 3
Received response	0.088** (1.99)	0.361** (2.10)
N	211424	7023

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 18: Change in the star rating of a consumer's second review, as function of the first review receiving a response.

	(1)
Received Response \times Second Review	0.567** (2.58)
Second Review	0.956*** (7.84)
Received Response	-0.183 (-0.35)
N	358
R ² within	0.42

Note: The dependent variable is the star of the current review.

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