Channels of Impact: User reviews when quality is dynamic and managers respond

Judith A. Chevalier
Yale School of Management and NBER

Yaniv Dover
Jerusalem School of Business, Hebrew University

and

Dina Mayzlin
University of Southern California, Marshall School of Business

This draft: August 2015

Preliminary and Incomplete - Do not post. Authors contributes equally to the manuscript, and the names are are listed in alphabetical order.

1 Introduction

Many Internet sites rely extensively on user contributions. User-generated online reviews have become an important resource for consumers making purchase decisions: retailers like Amazon rely on reviews to help match consumers with products, and information portals like Tripadvisor and Yelp have user-generated reviews at the core of their business models. One important similarity between Amazon and Tripadvisor is that both connect multiple buyers and sellers: Amazon helps customers pick a vacation guide book, and Tripadvisor helps customers choose a hotel. We refer to these connector sites as "platforms."

Even on platforms with similar purposes, there is a substantial amount of variation in the details of platform design. For example, some sites verify that reviewers purchased the product, while others do not. Some platforms allow reviewers to edit reviews after posting them; some do not. Despite the importance of online reviews to both consumers and firms, it is not clear how platform design features impact user review posting behavior. In part, this is because it is not clear why consumers are motivated to expend time and energy posting reviews. In this paper, we examine the impact of platform design features on user review posting behavior. A distinguishing feature of the environment we study is that the quality of the product being reviewed can evolve over time.

In an environment in which product quality is time-invariant, customer reviews are most naturally thought of as an avenue for customers to share information with each other. Firms may, of course, be an audience for reviews in that they may consider reviews in designing future products. However, given that the products being reviewed will not change as a function of the review, the owner or manager of the product is likely not considered by consumers to be the primary audience for the review. In this paper, we study a product category for which product quality is dynamic and managerial investments may alter product quality over time. In such cases, the consumer may reasonably view both the management and other consumers as an audience for the review.[1] In quality-variant product categories, a number of platforms have increased the salience of the management as an audience for reviews by allowing firms to directly respond to customer reviews. The

[1] There is a sense in which the quality of physical products such as books or beauty products could change over time. A book could become dated, or a products attributes could become dominated by a newly-introduced product. However, these changes are not a result of managerial investments as with hotels and restaurants and thus, the impact issues we discuss are not completely relevant.

entry of the firm into the conversation potentially changes the nature of the discourse, which in turn may impact the customers incentives to post reviews. That is, while the firm always had the ability to collect and analyze customer reviews, [2] the ability to respond allows the firm to credibly signal that it is reading reviews and responding to suggestions therein. The response functionality can transform a peer-to-peer review system into a hybrid peer review/customer complaint system.

We examine whether managerial response stimulates consumer reviewing and we examine whether it changes the nature of the reviews posted. Examining the relationship between the managerial response feature and consumer reviewing behavior yields insights with important implications for both managers and academics. Our research contributes to the nascent literature on reviewers' motivations. Understanding these motivations in turn may aid the platform hosts optimal design as well as the firms decision-making when responding to consumer comments.

Reviewing platforms that allow managerial response generally allow it in one of two ways. The first is private managerial response; owners or managers of reviewed businesses can send a private communication to the reviewer about the review.[3] The more common type of managerial response, and the only type allowed by most reviewing sites, is public; the manager cannot communicate privately and directly with the reviewer, but can post a public textual response to the reviewer. Interestingly, the reviewer only sees the reply if he or she happens to come back and look at the reviewed service again (though some sites email reviewers when responses are posted). However, other consumers and other potential reviewers will see the communication. Despite the fact that manager review responses are often written in the style of personal communications (e.g. "I am sorry to hear that everything was not to your satisfaction"), other consumers are in fact likely the primary audience for the managerial response.

We examine managerial response in the empirical setting of midtier to luxury hotels. The primary hypothesis that we examine is whether public managerial response communication stimulates reviewing activity. Quite

[2]The fact that consumer reviews contain managerially-useful information is illustrated by White (2014) which discusses the trend of hotels using customer reviews as "blueprints" for renovation.

[3]Yelp is a prominent example of a reviewing platform that allows private communication. Yelp also (unusually) allows reviewers to change their reviews and the private communication channel is often used to encourage reviewers to change a review. See, for example, the Yelp Official Blog, http://officialblog.yelp.com/2011/03/tactics-for-responding-to-online-critics-new-york-times-boss-blog.html.

simply, the idea is this: because someone is listening, consumers talk more. Specifically, we investigate whether the introduction of managerial responses lead consumers to write more reviews and whether it leads them to put more effort into reviewing, as measured by writing longer reviews. However, this simple idea leads to a slightly more subtle hypothesis about review valence.

One of the primary motivations for responding to a review may be to encourage consumers to view the experience described in a negative review as unlikely to be repeated. Indeed, much of the online discussion about managerial response strategies emphasizes the importance of responding to negative reviews and advice about how to respond to them [4] We will demonstrate that more negative reviews are more likely to receive managerial responses on reviewing platforms than are more positive reviews and that responses to negative reviews tend to be much longer (and therefore, presumably more substantive) than the responses to more positive reviews. However, in responding to reviews that contain criticisms, the manager signals to consumers that those criticisms will be read by management, possibly responded to, and possibly acted upon. If consumers write hotel reviews in order to have an impact on hotel quality, responding to reviews will encourage future reviewing activity. However, if consumers perceive that managers concentrate their response activity to negative reviews or if negative reviews receive responses that promise more action, consumers may be encouraged to post more negative reviews. Thus, while the motivation for responding to a negative review is to mitigate its effects, responding to a negative review may actually disproportionately encourage the posting of more negative reviews. While we have not seen others in the literature hypothesize that managerial responses stimulate negative reviewing, this hypothesis derives straightforwardly from extant evidence in the literature on reviewer behavior and the importance of negative reviews to management responders.

In this paper, we examine whether managerial response activity disproportionately stimulates negative review production. In particular, we study whether reviewing activity changes for a hotel following the first day of posting of managerial responses on three websites that allow managers to respond to reviews: TripAdvisor, Expedia, and Hotels.com.

Of course, in testing our hypotheses, we are faced with an identification challenge. Clearly, hotels that post managerial responses are different from hotels that do not. In time, the decision to commence posting responses

[4] See for example the many articles that TripAdvisor has written in its "TripAdvisor Insights" article series on the topic of managerial responses to negative reviews, http://www.tripadvisor.com/TripAdvisorInsights/t16/topic/management-responses.

is an endogenous decision of the manager. It is possible, for example, that managers begin responding to reviews when something is going on at the hotel that leads the manager to anticipate more reviews, and more negative reviews being posted in the future.

To handle this identification challenge, we employ an extension of the techniques we first employed to handle similar identification challenges in Chevalier and Mayzlin (2006) and in Mayzlin and Dover (2014). Specifically, our identification strategy has three components.

First, we undertake an "event study" technique in which we examine reviewing activity for a given hotel for very short (6 week) time windows before and after the posting of the first managerial response. Therefore, we are not identifying the impact of review responses by straightforwardly comparing hotels that do and do not post responses; we are examining the change for a given hotel over time. Furthermore, by examining only the before/after change surrounding a discrete event in a very short time window, we are discarding changes that may derive from longer run investments in facilities, position, or quality of the hotel. We examine whether reviewing activity changes in the very short time window for a given hotel before and after managerial responses are posted.

Second, we undertake specifications in which the reviewing changes on each of the focal sites on which managers may post responses are measured controlling for reviewing changes on other sites where managerial responses are not allowed. Specifically, managerial responses are not allowed on the popular booking sites Priceline and Orbitz, and our reviewing activity changes are measured from specifications which control for the changes in reviewing activity on Priceline and Orbitz. Thus, if a manager undertakes his/her first response on TripAdvisor in anticipation of some physical change in the hotel that will lead to negative reviews, the controls for changes in Priceline and Orbitz reviews should remove that source of review changes. This is similar to our approach in Chevalier and Mayzlin (2006) and Mayzlin and Dover (2014), though the appropriate "treatment" and "control" sites differ across the papers.

Finally, all of our specifications are conducted measuring the change in the hotel's reviews relative to changes in average reviews in the geographic area on the same reviewing site. For example, when we are measuring changes in TripAdvisor reviews around the commencement of managerial response, we are measuring the change in reviewing activity for the hotel relative to other hotels in the local area. This differencing strategy will prevent us from attributing the changes in reviewing activity on TripAdvisor to the managerial response by the hotel if, in fact, the change in reviewing activity was caused by, for example, increased TripAdvisor advertising in the local area. This is a variant of a strategy that we employed in Mayzlin and Dover (2014).

In sum, this "multiple-difference" strategy controls for many of the underlying sources of the identification challenge. We discuss more below the extent to which remaining confounding effects can be eliminated and our results can be interpreted as causal. We find that reviewing activity for a given hotel increases on TripAdvisor in the six week window following a managerial response relative to the six weeks prior and relative to the same hotel on Priceline and Orbitz, and relative to the TripAdvisor reviews of other hotels in the geographic area in the same six week window. We also find, relative to the controls, a statistically significant decrease in review valence and a statistically significant increase in the length of reviews. We have many fewer managerial response episodes for Expedia and Hotels.com. However, we conduct the same exercise for these hotels. For these hotels, as with TripAdvisor, we find an increase in reviewing activity following the posting of a managerial response. For Expedia, we also find a significant decrease in review valence, and for Hotels.com, a significant increase in review length.

We explore further the decrease in review valence using our large TripAdvisor sample. We divide our sample according to what kind of reviews the hotel responded to on its very first day of responding on TripAdvisor. Some hotels, on the first day of responding, responded only to negative reviews and some hotels responded to positive reviews or a mix of positive and negative reviews. We find that hotels that respond only to negative reviews have a significant increase in reviewing activity and a significant decrease in reviewing valence. Hotels that respond to higher-rated reviews also have an increase in reviewing activity, but have no significant change in reviewing valence. While we are cautious about interpreting this particular set of specifications causally, it does suggest that responding only to negative reviews could lead consumers to believe that negative reviews will receive managerial attention and be responded to, differentially increasing negative reviewing effort.

Our paper proceeds as follows. Section 1 reviews the literature. Section 2 describes our data and provides summary statistics about our sample. Section 3 describes our methodology. Section 4 provides our basic results. Section 5 discusses robustness issues. Section 6 concludes.

2 Related Literature

A growing literature has addressed the question of why users post reviews. For example, Berger (2014) suggests the following drivers of posting behavior: altruism, impact, impression management, emotion regulation and social bonding. Wu and Huberman (2010) provide evidence that individuals are more likely to post reviews on Amazon when their reviews will have a greater impact on the overall average review. That is, reviewers are more likely to post reviews when their opinion differs more from the consensus and when there are fewer reviews. Given the overall positive valence of reviews on review sites, the impact hypothesis is consistent with the findings of Moe and Trusov (2011). They examine data from a beauty products site and demonstrate that reviews become increasingly negative as ratings environments mature. **?** test the hypotheses of Wu and Huberman (2010) using book data from Amazon.com and show that low-impact reviews are more likely to be posted when reviewing costs are low. Specifically, they demonstrate that the sequential decline in ratings is more pronounced for high- than for low-cost reviews. [5]

Our paper is also related to recent literature on platform design and mechanisms for eliciting consumer feedback. Numerous papers document that many consumers who use an internet site do not post reviews. For example, **?** show that only thirteen percent of buyers posted a review on a European hotel booking portal. Nosko and Tadelis (2014) show that sixty percent of consumers leave feedback on Ebay while Fradkin and Pearson (2015) show that about two thirds of consumers leave reviews on Airbnb. The results of both Nosko and Tadelis (2014) and Fradkin and Pearson (2015) suggest that eliciting negative feedback from consumers may be particularly difficult. For example, Nosko and Tadelis (2014) suggests that consumers that have had negative experiences are less likely to leave feedback. Similarly, Horton and Golden (2015) demonstrates a positive skew in reviewing on the temporary-labor site ODesk.

There is only a limited nascent academic literature on managerial response to reviews. The papers of which we are aware are Park and Allen (2013); Ye and Chen (2009); Proserpio and Zervas (2014) and Kim and

[5] In a context distinct from product reviews, Zhang and Zhu (2011) provide support for the impact hypothesis using data from Chinese Wikipedia postings. Zhang and Zhu (2011) show that contributions to Chinese Wikipedia by contributors outside mainland China declined dramatically when the site was blocked to readers in mainland China. This suggests that contributors receive social benefits from posting their contributions; reducing the size of potential readership reduces the benefits of contributing.

Brymer (2015). Park and Allen (2013) use a case study of four luxury hotels; the authors examine why management choose to be active or not in review responses. Kim and Brymer (2015) use proprietary data from an international hotel chain and show a correlation between responses to negative comments in online reviews and hotel performance. Ye and Chen (2009) conduct an experiment similar to ours and Proserpio and Zervas (2014). Using data from two Chinese travel agents, Ye and Chen (2009) show that reviewing activity and valence increases for hotels that post manager responses on the travel site that allows responses relative to the travel site that does not. However, the number of manager responses is quite low.

The closest paper to ours is Proserpio and Zervas (2014). This paper examines Texas hotels, comparing the six month window before and after the initiation of managerial response on TripAdvisor. The identification scheme is very similar to ours; however, these authors use Expedia as a control for TripAdvisor, constraining their sample to hotels that do not use Expedia's response function. Interestingly, they have very different hypotheses than ours; they hypothesize that responses will improve valence by establishing reciprocity. They find evidence of valence improvement following the establishment of managerial response. The contrast between their results and ours may yield interesting insights.

3 Data and Methodology

The starting point of our data collection efforts is the identification of 50 focal cities. As in citetpromoreviews, we identified the 25th to 75th largest US cities to include in our sample. Our goal was to use cities that were large enough to have many hotels, but not so large and dense that competition patterns amongst hotels would be difficult to determine. Using data available on the Tripadvisor website in mid-2014, we identified all hotels that Tripadvisor identifies as operating in these focal cities. We also obtained data from Smith Travel Research, a market research firm that provides data to the hotel industry (www.str.com). STR attempts to cover the universe of hotels in the US. We use name and address matching to hand-match the TripAdvisor data to STR.

From STR, we obtain many characteristics of hotels including their class gradings of hotels. STR grades hotels by brand and objective characteristics into six quality tiers. For this paper, we exclude the bottom two quality tiers. The excluded chains consist largely of roadside budget motel chains such as Red Roof Inn, Super 8, Motel 6, Quality Inn, and La Quinta Inns.

We include hotels in STR's "Upper Midtier" range and higher. "Upper Midtier" includes Holiday Inn, Hampton Inn, Fairfield Inn, and Comfort Inn. We focus on "Upper Midtier" hotels and higher because "Economy Class" and "Midscale Class" hotels have significantly fewer reviews per hotel and hotel shoppers can perhaps be expected to be less quality-sensitive. Our sample of hotels is more homogeneous. Furthermore, the Revinate data that we will use for reviews has much better coverage for "Upper Midtier" hotels and higher. In total, we obtain a sample of 2104 hotels that match between STR and TripAdvisor in the "Upper Midtier" or higher categories.

Finally, our main source of review data comes from Revinate, a guest feedback and reputation management solutions provider. Among other services, Revinate provides client hotels a Review Reporting Dashboard. Client hotels can view daily social media feed from all major reviewing sites on one page, view the equivalent feed for their competitors, and respond to reviews from multiple sites from the single interface. In order to provide this service, Revinate has invested in creating robust matching of the same hotel across multiple reviewing platforms.

Revinate has excellent coverage of the 2104 hotels in our target sample. Revinate has substantial market share; they serve more than 23000 client hotels worldwide. Many large chains subscribe to Revinate services for all of their hotels. Crucially for us, they track not only their clients, but a large group of client competitors. Overall, of the 2104 hotels in our target sample, we are able to obtain Revinate data for 88 percent of them, for a total of 1843 hotels.

Even with this excellent coverage, the imperfect coverage presents a selection bias. However, note that the hotels that we track contain both Revinate clients and non-clients. The 261 hotels that we could not track through Revinate are clearly not Revinate clients. In the analysis that follows, we will undertake weighted specifications in which the non-clients receive more weight to equate the weight on non-clients in the sample to the weight of non-clients in the overall population.

The Revinate data contains full text review information with date and time stamps for every review for the sites that Revinate tracks for our 1843 hotels. They also collect full text of all managerial responses, again with date and time stamps. Due to a peculiarity of the way that Booking.com displays review data, our review data for Booking.com does not contain older reviews for the site. We include some summary information about Booking.com in this section, but do not use Booking data at all in our analyses.

Table 1 contains summary statistics that describe the reviewing and response data for our sample of hotels for six of the most popular booking and reviewing sites: Booking.com, Expedia, Hotels.com, Orbitz, Priceline, and TripAdvisor.

Table 1 reveals several interesting stylized facts about the data. First, note that there are no review responses on Booking, Priceline, and Orbitz because they do not allow responses. TripAdvisor, Expedia, and Hotels.com allow responses. In our sample, roughly half of TripAdvisor reviews receive responses while only about 8 percent of Expedia reviews receive responses and only 2 percent of Hotels reviews receive responses.

Revinate customers are much more likely to respond to reviews than non-customers. This is likely partly due to selection; they have demonstrated their interest in social meeting management by becoming customers. However, this is also likely due to the causal impact of the Revinate platform. The platform makes it very easy to respond to reviews. Notice that the customer vs. noncustomer review response rates are most disparate for non-TripAdvisor sites. This makes sense because the Revinate interface makes it equally easy to view all of the sites and post all of the responses; non-Revinate customers may have more tendency to monitor only the perceived most influential site, TripAdvisor. Again, our reweighting of customers vs. non-customers is important to achieve representativeness.

Important for our purposes is the disparity between the valence of reviews that are responded to and reviews that are not responded to. Reviews that are responded to average 4.0 for TripAdvisor versus 4.2 for reviews that do not receive responses. This disparity is greater for the other sites.

The differences in the valence of reviews responded to versus not responded to is largely due to the reviews selected for response rather than the characteristics of the hotels that respond to reviews. Limiting the sample of hotels to only hotels that have responded to five reviews produces similar summary statistics. In unreported specifications, for each of the three sites that allow responses, we take the sample of all reviews as observations, and regress the indicator variable for "manager responded" on the review rating and hotel fixed effect. The coefficient for the review rating is strongly negative and significant and is larger in magnitude and significance than the coefficient when fixed effects are not included. This suggests that better hotels respond to their worse reviews.

Table 2 provides summary data on review length of each of the sites. Clearly, there are differences across sites in typical review length. For example, reviews on Booking.com are particularly short and reviews on TripAdvisor.com are particularly long. Worse reviews tend to be significantly more detailed across all sites.

Table 2 also provides summary data on review response length for the

Table 1: Summary Statistics - Reviews

	Booking	Expedia	Hotels.com	Orbitz	Priceline	TripAdvisor
Number of reviews						
noncustomer	248065	252781	223450	94321	256143	624555
customer	80596	106207	81122	36973	93841	265944
weighted	337,465	367,960	312,503	134,642	359,075	912,666
Response share noncustomer	0	0.056	0.015	0	0	0.457
customer	0	0.132	0.041	0	0	0.517
weighted	0	0.078	0.022	0	0	0.474
Mean Rating						
noncustomer no response	4.142	4.165	4.237	3.892	3.988	4.139
customer noresponse	4.146	4.158	4.241	3.891	4.011	4.194
weighted no response	4.143	4.163	4.238	3.892	3.994	4.155
noncustomer response		3.963	4.1309			4.0259
customer response		3.950	4.015			4.102
weighted response	0.000	3.959	4.101	0.000	0.000	4.048
Weighted: response vs no response						
Percentage difference		-4.9%	-3.2%			-2.6%
For the universe of hotels that have responded 5 times before the data of this review:						
Mean Rating						
noncustomer no response	4.142	4.163	4.234	3.887	3.989	4.135
customer no response	4.143	4.156	4.238	3.888	4.012	4.190
weighted no response	4.143	4.161	4.235	3.887	3.995	4.151
noncustomer response		3.964	4.131			4.026
customer response		3.949	4.015			4.102
weighted response		3.960	4.101			4.048
Weighted: response vs no response						
Percentage difference		-4.8%	-3.2%			-2.5%

Table 2: Review length and managerial response length

	Booking	Expedia	Hotels.com	Orbitz	Priceline	Tripadvisor
Review stars $\leq$ 3 review length	221.291	473.814	359.919	464.867	324.862	964.677
Review stars>3 review length	138.711	362.464	229.833	347.809	231.359	683.175
Percentage difference	-37.3%	-23.5%	-36.1%	-25.2%	-28.8%	-29.2%
Review stars $\leq$ 3 response length		518.891	514.179			703.007
Review stars>3 response length		387.266	361.037			516.219
Percentage difference		-25.4%	-29.8%			-26.6%

All means are weighted to overrepresent Revinate non-customers.

three sites that allow review responses. Managers tend to provide longer review responses on TripAdvisor versus the other two sites. Across all sites, managers appear to put more effort into responding to negative reviews. The summary table shows that, for all three sites that allow responses, responses to more positive reviews (greater than three stars) are 25 to 30 percent shorter than responses to more negative reviews.

4 Methodology

As discussed above, we consider the introduction of managerial responses on a particular site to potentially present a discrete change in site experience for customers who visit the site. Thus, our methodology focuses on changes in reviewing activity in a very tight time window around the day that responses are first posted.

In Table 3, we examine summary statistics on the behavior of hotels on the day of first managerial response posting. The first thing to notice about Table 3 is that managers frequently respond to more than one review the first time that review responses are posted. The average star of reviews responded to, unsurprisingly, are more negative than the overall population of reviews. For example for TripAdvisor, 641 of the 1807 hotels that post

Table 3: Characteristics of first day of responses

	Expedia	Hotels.com	Tripadvisor
Reviews responded to the first day			
Average star Noncustomer	3.54	3.78	3.12
Average star Customer	3.65	3.56	3.20
Average star weighted	3.57	3.72	3.14
Total number responded Noncustomer	1.91	1.69	1.94
Total number responded Customer	1.99	1.32	2.04
Number responded to weighted	1.93	1.60	1.97

responses, responses are posted to reviews with an overall average star rating of strictly less than three on the first day of response posting.

Our identification scheme relies on a differences methodology. Our primary specifications examine the 6 week window before and after the posting of first managerial response. We undertake robustness specifications below, but describe our primary measurement strategy here.

We will describe the platforms in which the manager initially posts a response as the "treatment" platforms. These are TripAdvisor, Expedia, and Hotels.com which we will consider separately in separate specifications (since the response posting time windows are different for the three sites). The "control" platforms are Orbitz and Priceline.

First, in order to control for contemporaneous factors that may cause within-platform changes in reviews in a geographic area, we straightforwardly difference the before vs. after review measurements from the before vs. after review measurements for the geographic area. We construct the geographic mean for each TripAdvisor geocode imposing no restrictons on the hotels included in the mean calculation. Our measure of review changes measures the over time of the difference from the geographic mean for both treatment and control platforms. Thus, for example, in all specifications that use the number of TripAdvisor reviews in a time window, TripAdvisor reviews for the observation hotel is calculated as the difference between TripAdvisor reviews in the time window minus the average number of reviews for all other hotels in the same city for the same time window.

We have two platforms that we use as controls, Orbitz and Priceline.

An important issue in using one platform as a control for another is that the platforms certainly may cater to different populations. This might be particularly true for TripAdvisor vs. the other platforms, as TripAdvisor is primarily a review platform while Expedia, Orbitz, and Priceline are booking platforms. In including two controls, we allow the data to "choose" the combination of platforms that are the best control. Thus, for any review measurement constructed for our treatment sites TripAdvisor, Expedia, and Hotels, we construct the corresponding measure for both Orbitz and Priceline and use these as control variables in our regression specifications.

Table 4 summarizes the cross-sectional correlation in number of reviews, review valence, and review length across sites but within hotels. Our identification strategy is predicated on the idea that hotel characteristics will be similarly measured across sites. Table 4 suggests high correlation among the sites for all of the measures. The correlation across sites is largest for review valence and smallest for review length. Hotels that inspire longer reviews on one site tend to receive longer reviews on other sites, but the effect is modest in magnitude. Our two control sits, Orbitz and Priceline are less correlated with each other across each of the measures than are any other pair of sites. This suggests that there is plausibly different information about the hotel captured by using each of them as a control.

Recall that our review measures of interest are: change in the number of reviews between the six week windows, change in the valence of reviews, and changes in measures of the length of reviews. For each measure, M, for hotel i in city j, our simple estimating equation is :

$$\begin{aligned}
(M_{ij}^{Treat} - \overline{M}_j^{Treat})_{Post} - (M_{ij}^{Treat} - \overline{M}_j^{Treat})_{Pre} \quad &= \\
\alpha + [(M_{ij}^{Price} - \overline{M}_j^{Price})_{Post} - (M_{ij}^{Price} - \overline{M}_j^{Price})_{Pre}]\beta_1 &+ \\
[(M_{ij}^{Orbitz} - \overline{M}_j^{Orbitz})_{Post} - (M_{ij}^{Orbitz} - \overline{M}_j^{Orbitz})_{Pre}]\beta_2 &+ \epsilon_{ij}
\end{aligned} \tag{1}$$

In this equation, "Pre" denotes the six week period prior to first managerial response on the treatment site for hotel i, "Post" denotes the six week period following the first managerial response on the treatment site. The treatment site, Treat$\in$(TripAdvisor,Expedia,Hotels.com) and $\overline{M}_j$ denotes the city-average variable. Note that, for each specification, α is our variable of interest, as it is the change in the variable net of the changes in the controls.

Our identification strategy is helpful in overcoming several potential endogeneity challenges. When will our strategy fail? The primary weakness of

Table 4: Correlations within hotel and across sites

Review count correlations					
	TripAdvisor	Expedia	Hotels	Priceline	Orbitz
TripAdvisor	1.00				
Expedia	0.78	1.00			
Hotels.com	0.63	0.58	1.00		
Priceline	0.38	0.34	0.31	1.00	
Orbitz	0.67	0.41	0.29	0.24	1.00
Average star correlations					
	TripAdvisor	Expedia	Hotels	Priceline	Orbitz
TripAdvisor	1.00				
Expedia	0.81	1.00			
Hotels.com	0.71	0.80	1.00		
Priceline	0.72	0.77	0.72	1.00	
Orbitz	0.63	0.66	0.60	0.59	1.00
Length correlations					
	TripAdvisor	Expedia	Hotels	Priceline	Orbitz
TripAdvisor	1.00				
Expedia	0.55	1.00			
Hotels.com	0.45	0.58	1.00		
Priceline	0.30	0.34	0.31	1.00	
Orbitz	0.34	0.41	0.29	0.24	1.00

our strategy is that it is possible that a hotel-specific time-specific platform-specific factor is correlated with both the initiation of managerial response for hotel i and with the future review process for hotel i on that specific platform. We find the possibility of one category of such confounds slightly more plausible for Expedia and Hotels.com than for TripAdvisor.com. This is because Expedia and Hotels.com are both booking sites, rather than purely a review site. It is possible, for example that hotels systematically simultaneously commence participation in hotel-specific promotions on Expedia or Hotels.com and initiate managerial response on that site. Such a promotion might plausibly lead, through an increase in platform-specific bookings, to an increase in reviewing activity on those sites. Even if a hotel undertook some kind of promotion on TripAdvisor, the consumer would click out to a different site to book the reservation. While this is a concern, we note three things about this concern: First, many of the plausible confounds that we can think of (such as promotions) might lead people to book more rooms in the short run, but the booking activity should somewhat more slowly work its way into staying and reviewing activity (since many people book for a stay occurring somewhat in the future). Second, it is not entirely clear why such a promotion would also lead people to be systematically differentially satisfied or unsatisfied with their experience; that is, there is not a natural prediction for review valence. Third, there is also not a natural prediction for review length.

A second scenario that would undermine our strategy is the possibility that hotels that commence review response on a site are "getting organized" about catering to the users of the site more generally. This might be a particular concern for TripAdvisor, since TripAdvisor has become so important in the industry. Several factors, we believe, mitigate this concern in our setting. First, this is somewhat less likely to be an important issue for Expedia and Hotels.com since TripAdvisor is so important in the industry. Second, the natural bias stories of this type seem to cut in the opposite direction of our hypothesis and findings. If a hotel were launching a coherent TripAdvisor management system with immediate implications, it is hard to see how that would systematically lead to a decrease in review valence. Third, there is not a clear natural prediction of this possibility for review length.

5 Results

Table 5 presents summary statistics of the variables used to estimate Equation 1.

Table 5: Summary statistics for six week difference variables

Summary statistics for first TripAdvisor response	Number of Reviews	Average Stars	Average Length
TripAdvisor average	4.50	4.01	762.60
(sum six weeks before)	(8.29)	(0.82)	(412.21)
Diff in Diff TA	1.27	-0.06	50.61
	(5.64)	(0.94)	(514.00)
Diff in Diff Orbitz	0.14	0.00	6.54
	(2.96)	(2.35)	(344.55)
Diff in Diff Priceline	0.12	0.05	3.56
	(2.78)	(2.28)	(174.78)

Summary statistics for first Expedia response	Number of Reviews	Average Stars	Average Length
Expedia average	4.97	4.19	351.13
(sum six weeks before)	(6.55)	(0.66)	(191.49)
Diff in Diff Expedia	0.33	-0.07	9.94
	(4.59)	(0.81)	(239.75)
Diff in Diff Orbitz	-0.04	-0.13	-5.76
	(2.20)	-(3.34)	(364.92)
Diff in Diff Priceline	0.12	0.07	16.59
	(3.64)	(2.20)	(255.96)

Summary statistics for first Hotels.com response	Number of Reviews	Average Stars	Average Length
Hotels.com	9.49	4.25	236.77
(sum six weeks before)	(13.19)	(0.56)	(132.19)
Diff in Diff Hotels	1.06	0.00	24.23
	(8.57)	(0.61)	(207.88)
Diff in Diff Orbitz	-0.04	0.30	-1.19
	(3.61)	(2.10)	(337.62)
Diff in Diff Priceline	-0.26	0.19	8.04
	(4.24)	(2.16)	(192.40)

Standard deviations are in parentheses. Observations are weighted using the probability weight for Revinate customers vs noncustomers.

Table 6 provides estimates of Equation 1 where TripAdvisor is the treatment site. Column 1 examines the variable of primary interest to us, the change in the number of reviews for the sample of 1807 first responders on TripAdvisor. Note that both the Orbitz control and the Priceline control coefficients are positive and significant at least at the ten percent level. The number of reviews positively comoves across the sites. This is not surprising, as presumably all of the sites get more reviews in periods of particularly high occupancy for the hotel relative to the local area. Note that both Orbitz reviews and Priceline reviews also experience review increases over the period. The constant term is positive and statistically significant at the one percent level. This suggests that, net of the controls, the number of reviews increase by 1.2 reviews. The average number of reviews in the pre-review period is 4.5 (from Table 6), so this represents a substantial sudden increase in reviewing activity relative to the controls.

Table 6: TripAdvisor changes in reviewing activity

VARIABLES	(1) Num Reviews	(2) Avg Star	(3) Avg Length	(4) Length 1/2/3 star reviews	(5) Length 4/5 star reviews
Orbitz controls	0.117**	0.001	-0.015	0.039	0.007
	(0.059)	(0.011)	(0.048)	(0.079)	(0.048)
Priceline controls	0.110*	-0.003	-0.020	0.104	0.089
	(0.061)	(0.013)	(0.076)	(0.224)	(0.074)
Constant	1.240***	-0.064**	50.783***	19.492	31.235**
	(0.130)	(0.027)	(14.781)	(32.402)	(15.548)
Observations	1,807	1,210	1,210	516	1,037

Robust standard errors in parentheses
Regressions weighted to overrepresent Revinate non-customers
*** $p<0.01$, ** $p<0.05$, * $p<0.1$

Column 2 examines average stars. When considering average stars, we restrict the sample to hotels that had reviews in both the pre- and post- six week windows. We see that average stars decrease significantly for hotels that have commenced managerial response. This is consistent with our hypothesis that reviewers respond with more substantive reviews when they receive feedback that managers are listening. Interestingly, our results contrast Proserpio and Zervas (2014) in this regard. The point estimate is -0.06. This may seem small, but recall that the average stars obtained in the pre-period are 4.01 with a cross-sectional standard deviation of only 0.8. Thus, small movements in the average stars can significantly move a hotel up or down the rank ordering of hotels in a local area. An important issue consider is that the Orbitz and Priceline controls are not statistically significant in the regression. The hotels also have essentially no average star movement of Priceline or Orbitz over the two six week windows. This is to be expected if underlying quality at the hotel is, indeed, not significantly changing.

Column 3 displays results for the average length of review, with Columns 4-5 demonstrating review length for reviews of different star levels. Again, we restrict the sample to hotels that had reviews in both the pre- and post- period. Again, review length has not changed significantly over the 6 week windows for Priceline and Orbitz, suggesting that perhaps nothing has changed at the hotel that fundamentally leads consumers to want to leave longer reviews. However, reviews on TripAdvisor increase significantly. The point estimate of the increase is 51 characters. Again, this is a large change given the average length of reviews on TripAdvisor for the period prior to the managerial response is 763 characters. We examine the change in review length for negative valence (1,2, and 3 star) reviews and positive valence (4 and 5 star) reviews separately. In each case, we must restrict the sample to hotels that have reviews in that category in both the pre and post windows. This increase in reviewing effort is estimated to be positive both for negative and positive valence reviews, although is is statistically different from zero only for the 4 and 5 star reviews.

Table 7 provides estimates of Equation 1 where Expedia is the treatment site. Of course, the sample of review responses is much smaller. Interestingly, the number of reviews that the hotel has in the six weeks prior to the first response is about the same for Expedia as for TripAdvisor, suggesting that responders on Expedia tend to be hotels that garner a lot of Expedia reviews. Despite the smaller number of observations, we find somewhat similar, albeit weaker, results. The increase in the number of reviews net of the controls is 0.3, about one-quarter of the magnitude of the effect estimated for TripAdvisor, but still significantly different from zero at standard confidence level. The decrease in review valence of 0.07 is very similar to the TripAdvisor results. The increase in review length is small and insignificant, however.

Table 8 provides estimates of Equation 1 where Hotels.com is the treatment site. Here, the sample of responding hotels is again quite small, as we have only 332 Hotels that provide responses on Hotels.com. Nonetheless, we do find a significant increase in the number of reviews posted, no measurable change in average star, a small (and significant at the ten percent level) increase in review length overall.

Overall, we interpret the results in Tables 6, 7, and 8 as demonstrating that reviewing activity increases following managerial response, that valence decreases (at least modestly), and that the length of reviews (a measure of reviewing effort) increases.

Table 7: Expedia changes in reviewing activity

	(1)	(2)	(3)	(4)	(5)
VARIABLES	Num Reviews	Avg Star	Avg Length	Length 1/2/3 Star reviews	Length 4/5 Star reviews
Orbitz controls	0.135	0.006	0.026	-0.051	-0.005
	(0.086)	(0.012)	(0.024)	(0.054)	(0.032)
Priceline controls	0.086*	0.017	-0.029	0.007	-0.017
	(0.052)	(0.013)	(0.046)	(0.077)	(0.053)
Constant	0.329**	-0.072**	10.568	-30.860	9.557
	(0.145)	(0.030)	(9.055)	(24.113)	(9.056)
Observations	984	709	709	275	652

Robust standard errors in parentheses
Regressions weighted to overrepresent Revinate non-customers
*** p<0.01, ** p<0.05, * p<0.1

Using our large sample of TripAdvisor first responses, we consider hotels first-day reviewing strategy. In doing so, we caution that we do not know why some hotels pursue different first-day reviewing strategies and that, as discussed above, the weakness of our identification strategy stems from circumstances in which there are review creation strategies that are simultaneously platform- and hotel- and time-specific. Nonetheless, since our hypothesis is that customers will more readily initiate complaints on TripAdvisor when they believe that complaints will be listened to and acted upon, we examine whether managerial response to complaining differentially

Table 8: Hotels.com changes in reviewing activity

VARIABLES	(1) Num Reviews	(2) Avg Star	(3) Avg Length	(4) Length 1/2/3 Star Reviews	(5) Length 4/5 Star Reviews
Orbitz controls	0.412***	0.004	0.028	-0.036	-0.012
	(0.132)	(0.019)	(0.038)	(0.077)	(0.038)
Priceline controls	0.159	-0.022	-0.032	0.128	-0.026
	(0.135)	(0.019)	(0.055)	(0.096)	(0.065)
					(0.065)
Constant	1.116**	0.001	24.515*	12.047	19.924
	(0.460)	(0.038)	(12.835)	(30.852)	(13.860)
Observations	332	265	265	136	256

Robust standard errors in parentheses
Regressions weighted to overrepresent Revinate non-customers
*** p<0.01, ** p<0.05, * p<0.1

stimulates negative reviews on the part of reviewers.

In Tables 9 and 10, we reestimate Equation 1 for TripAdvisor but separate the data into two separate groups. First, in Table 9, we consider the set of TripAdvisor respondees whose first day responses are to reviews with an average star value of less than three. Second, in Table 10, we consider the disjoint set of TripAdvisor respondees whose first day responses are to reviews with an average star value of three or more. For the 641 hotels that respond to low average star reviews in Table 9, we see that, in the six weeks after the first response, the number of reviews increases significantly, review valence decreases significantly and average length increases. The decrease in review valence is roughly double the magnitude of our estimate of the review valence differences using the overall sample. On closer inspection of the pattern of review length increases, review length increases are positive but insignificant for the two valence categories. Overall, the results in this table are suggestive that managerial responses that concentrate on responding to negative reviews encourages more and more detailed negative reviews from consumers.

For the 839 hotels that respond to higher average star reviews in Table 10, we see that, in the six weeks after the first response, the number of

Table 9: Changes in reviewing activity for TripAdvisor reviewers who respond to bad reviews

VARIABLES	(1) Num Reviews	(2) Avg Star	(3) Avg Length	(4) Length 1/2/3 Star Reviews	(5) Length 4/5 Star Reviews
Orbitz controls	0.093	-0.004	-0.046	0.085	0.002
	(0.071)	(0.016)	(0.075)	(0.112)	(0.077)
Priceline Controls	-0.035	-0.000	-0.041	-0.091	0.107
	(0.055)	(0.017)	(0.108)	(0.225)	(0.107)
Constant	0.724***	-0.143***	62.186***	42.942	25.324
	(0.157)	(0.037)	(20.537)	(41.472)	(20.336)
Observations	968	673	673	305	564

Robust standard errors in parentheses
Regressions weighted to overrepresent Revinate non-customers
*** p<0.01, ** p<0.05, * p<0.1

reviews increases significantly, review valences are unchanged (the point estimate is positive) and average review length increases modestly. The contrast between Table 9 and Table 10 is suggestive that negative reviewing activity is differentially stimulated when managers use the response function exclusively to respond to negative reviews. Again, this makes sense in an environment in which consumers post reviews in order to have an impact on hotel quality.

6 Robustness

We undertake a number of robustness specifications; for the robustness specifications, we focus on TripAdvisor, as we have a larger number of responders on TripAdvisor.

First, we investigate using a longer window both before and after the managerial response. In Table 11, we undertake the basic specifications of Equation 1 for TripAdvisor, allowing a 10 week window before and after the managerial response. The cost of a longer window is that it is more plausible that long-run investments in hotel quality that could be systematically coincident with the advent of managerial response are being experienced by

Table 10: Changes in reviewing activity for TripAdvisor reviewers who respond to good reviews

VARIABLES	(1) Num Reviews	(2) Avg Star	(3) Avg Length	(4) Length 1/2/3 Star Reviews	(5) Length 4/5 Star Reviews
Orbitz controls	0.161	0.007	0.012	-0.008	0.012
	(0.104)	(0.016)	(0.061)	(0.120)	(0.061)
Priceline controls	0.269**	-0.009	-0.005	0.348	0.071
	(0.113)	(0.018)	(0.108)	(0.402)	(0.105)
Constant	1.806***	0.034	37.490*	-9.838	38.204
	(0.210)	(0.040)	(20.990)	(50.474)	(24.044)
Observations	839	537	537	211	473

Robust standard errors in parentheses
Regressions weighted to overrepresent Revinate non-customers
*** p<0.01, ** p<0.05, * p<0.1

consumers. The benefit of a longer window is that, of course, over short time periods reviews and the correlation of the reviews across sites will be noiser.

The results in Table 11 look very similar to our base results in Table 6. The constant term in the number of reviews specification has increased from 1.24 to 2.27. Recall that we are measuring changes in reviews in levels; if the treatment effect of managerial response is permanent, we would anticipate roughly a 67 percent increase in the coefficient given the 67 percent increase in the time period. The average star measure is nearly identical in Table 11 and Table 6 as are the review length results.

We also examine potentially heterogeneity of results across hotel types. Recall that STR classifies all hotels into tiers using largely time-invariant characteristics. The tiers that we use in this paper are Luxury, Upper Upscale, Upscale, and Upper Midscale. The categories that we excluded are the “motel” categories Economy and Midscale. Even among the tiers we use, there is considerable heterogeneity in the type of hotel and the characteristics of the customers. The lowest tier we use, Upper Midscale, includes ordinary traveler hotels such as Hampton Inns or Fairfield Inns. The Luxury

Table 11: TripAdvisor specifications- longer time window

VARIABLES	(1) Num Reviews	(2) Avg Star	(3) Avg length	(4) Length 1/2/3 Star reviews	(5) Length 4/5 Star reviews
Orbitz controls	0.108*	0.008	-0.034	0.059	0.039
	(0.055)	(0.010)	(0.047)	(0.061)	(0.043)
Priceline controls	0.205*	0.005	0.015	0.052	0.024
	(0.105)	(0.011)	(0.063)	(0.121)	(0.061)
Constant	2.270***	-0.059**	35.761***	14.992	29.455**
	(0.209)	(0.023)	(13.188)	(26.845)	(12.769)
Observations	1,807	1,385	1,385	738	1,248

Robust standard errors in parentheses
Regressions weighted to overrepresent Revinate non-customers
*** p<0.01, ** p<0.05, * p<0.1

category includes hotels such as the Four Seasons and the Ritz-Carlton. It is reasonable to expect that reviewing dynamics and the role of managerial responses will differ across those hotels.

We investigate this in Table 12. In Table 12, we replace the constant term in each specification of Equation 1 with indicator variables for hotel type. Of the 1807 responding hotels, 560 are Upper Midscale, 635 are Upscale, 435 are Upper Upscale, and 149 are Luxury. Eventual responders represent 70 percent of the Upper Midscale hotels in our sample and over 90 percent of the other classes of hotels.

The results suggest that the change in reviewing activity is monotonically increasing in the ex ante hotel quality. That is, the increment to reviews is greatest for luxury hotels. However, the increases in review length are monotonically decreasing in the quality tier of hotels. That is, review length is particularly stimulated for Upper Midscale Hotels. The results for valence are mixed as large significant decreases in review valence are only found for the Luxury and Upscale hotels.

Additional robustness results to come

Table 12: Heterogeneity in review responses across hotel classes

VARIABLES	(1) Num Reviews	(2) Avg Star	(3) Avg length	(4) Length 1/2/3 Star reviews	(5) Length 4/5 Star reviews
Priceline controls	0.097	-0.004	-0.816	0.037	0.006
	(0.060)	(0.013)	(5.880)	(0.079)	(0.048)
Orbitz controls	0.109*	0.001	-0.014	0.107	0.101
	(0.057)	(0.011)	(0.075)	(0.225)	(0.074)
Luxury class	2.143***	-0.129**	3.767	32.848	-63.935
	(0.556)	(0.057)	(43.256)	(102.777)	(40.269)
Upper Upscale Class	1.987***	0.013	31.283	-7.630	15.282
	(0.347)	(0.044)	(27.520)	(46.301)	(31.625)
Upscale Class	0.965***	-0.140***	56.468**	81.052	29.319
	(0.186)	(0.050)	(25.789)	(70.591)	(26.284)
Upper Midscale Class	0.788***	-0.040	78.070***	-7.604	89.169***
	(0.184)	(0.057)	(28.078)	(62.779)	(28.776)
Observations	1,807	1,210	1,210	516	1,037

Robust standard errors in parentheses
Regressions weighted to overrepresent Revinate non-customers
*** p<0.01, ** p<0.05, * p<0.1

7 Conclusion

Allowing management to respond to user reviews has become a common feature of reviewing platforms, especially platforms geared to services such as hotels. We argue that allowing managerial response can fundamentally change the nature of the reviewing platform if users view themselves as in a dialogue with management rather than only leaving information for future customers. The nature of this communication is interesting because platforms typically require the manager's response to be public. How the manager responds could impact the consumer purchase decision in ways that we do not yet understand.

Casual empiricism and the advice proffered to hotel managers on the web suggests that managers attempt to use the response function to mitigate the impact of criticism, often by promising change. However, the prior literature suggests that consumers post reviews to have an "impact" on others and that they are more motivated to post when they perceive that they can have more impact. These observations lead to our hypothesis that the managerial response function promotes reviewing generally and promotes the production of critical reviews specifically. Our empirical results rely on a multiple-difference strategy to address endogeneity issues. Our results are generally supportive of the hypothesis that managerial response encourages critical reviewing.

These results have implications for the growing literature on online collaborative information creation. Our results suggest that seemingly small tweaks to the platform design can have measurable implications for consumer reviewing behavior and potentially the utility of reviews for future consumers. In the future, we intend to explore further why reviewing behavior may change when managerial response is implemented. In particular, we have begun exploring linguistic characteristics of reviews before and after the implementation of managerial response.

References

Berger, Jonah (2014) "Word-of-Mouth and Interpersonal Communication: An Organizing Framework and Directions for Future Research," *Journal of Consumer Psychology*.

Chevalier, J. and D. Mayzlin (2006) "The Effect of Word of Mouth on Sales: Online Book Reviews," *Journal of Marketing Research*, Vol. 43, pp. 345–354.

Fradkin, Grewal E. Holtz D., A. and M. Pearson (2015) "Bias and Reciprocity in Online Reviews: Evidence from Field Experiments on Airbnb," MIT Sloan School of Management and Airbnb working paper.

Horton, John J. and Joseph M. Golden (2015) "Reputation Inlfation in an Online Marketplace," NYU Stern working paper.

Kim, Lim Hyunjung, Woo Gon and Robert A. Brymer (2015) "The effectiveness of managing social media on hotel performance," *International Journal of Hospitality Management*.

Mayzlin, Chevalier Judith, Dina and Yaniv Dover (2014) "Promotional Reviews: An Empirical Investigation of Online Review Manipulation," *American Economic Review*.

Nosko, C. and S. Tadelis (2014) "The Limits of Reputation in Platform Markets: An Empirical Analysis and Field Experiment," University of California at Berkeley Working Paper.

Park, Sun-Young and Jonathan P. Allen (2013) "Responding to Online Reviews: Problem Solving and Engagement in Hotels," *Cornell Hospitality Quarterly*.

Proserpio, Davide and Georgios Zervas (2014) "Online reputation management: Estimating the impact of management responses on consumer reviews," Boston U. School of Management Research Paper.

White, Martha C. (2014) "Hotels Use Online Reviews as Blueprint for Renovations," *New York Times*, Vol. Sept 22.

Ye, Gu Bin, Qiang and Wei Chen (2009) "Measuring the Influence of Managerial Responses on Subsequent Online Customer Reviews- A Natural Experiment of Two Online Travel Agencies," Working Paper.

Zhang, Xiaoquan (Michael) and Feng Zhu (2011) "Group Size and Incentives to Contribute: A Natural Experiment at Chinese Wikipedia," *American Economic Review*.