Reciprocity in Two-sided Reputation Systems: Evidence from an Experiment on Airbnb[*]

Andrey Fradkin[†1], Elena Grewal[‡2], and David Holtz[§3]

[1]Boston University and MIT Initiative on the Digital Economy
[2]Unaffiliated
[3]MIT Sloan School of Management

March 30, 2020

Abstract

Reputation systems are used by nearly every digital marketplace but designs vary. One design choice is when to reveal post-transaction feedback to users, which may affect the content of reviews and the speed with which information is distributed in the marketplace. We study the effects of changing feedback revelation timing in a large-scale field experiment on Airbnb. Feedback in the treatment group was hidden until both parties submitted a review and was revealed immediately after submission in the control group. We show that the treatment reduced retaliation and reciprocation in feedback giving. The treatment also stimulated more reviewing in total, due to users' curiosity about what the counterparty wrote and/or the desire to have feedback visible to other users. The effects on feedback giving did not translate into a reduction of adverse selection as measured by subsequent transactions and prices.

[*] We are grateful to Chris Dellarocas, Liran Einav, Dean Eckles, Chiara Farronato, Shane Greenstein, John Horton, Caroline Hoxby, Ramesh Johari, Max Kasy, Jon Levin, Mike Luca, Jeff Naeckfer, Fred Panier, Catherine Tucker, Giorgos Zervas, and seminar participants at Microsoft, MIT, eBay, HKU, ACM EC'15, NBER Summer Institute, CODE, and the Advances in Field Experiments Conference for comments. We thank Matthew Pearson for early conversations regarding this project. The views expressed in this paper are solely the authors' and do not necessarily reflect the views of Airbnb, Inc. Fradkin, Holtz and Grewal were employed by Airbnb, Inc. for part of the time that this paper was written.

[†] Primary Author: fradkin@bu.edu

[‡] Primary Experiment Designer

[§] dholtz@mit.edu

1 Introduction

Reputation systems are used by nearly every digital marketplace to reduce problems stemming from information asymmetry and moral hazard. They do so by soliciting information about transaction quality and displaying it to other market participants. However, the creation of accurate reviews by market participants is voluntary and costly. As a result, reviews are under-provided in equilibrium in the absence of an appropriate compensation scheme (Avery, Resnick and Zeckhauser (1999)). This leads to missing information and a variety of biases, which can affect outcomes for both buyers and sellers. For instance, prior work shows that buyers who transact with sellers with biased reviews on a platform are less likely to transact on that platform in the future (Nosko and Tadelis (2015)). These factors make the design of effective reputation systems important for digital platforms.

We study the effects of an experimental intervention on the reputation system of Airbnb. Airbnb's reputation system is two-sided, meaning that the the guest and the host each have the opportunity to review one another. In the control group of our experiment, reviews are revealed to the counterparty as soon as they are submitted and users have up to 14 days to review. This leaves open the possibility that the second reviewer reciprocates or retaliates against the first review. Prior research has suggested that this type of behavior occurs on eBay and other platforms with bilateral review systems (Bolton, Greiner and Ockenfels (2012); Cabral and Hortaçsu (2010)). Our treatment changes the timing of when a review is revealed to the counterparty and on the platform. Reviews in the treatment are hidden until both parties have reviewed or until the time to review (14 days) has expired. After reviews are revealed, they cannot be modified, which makes it impossible to retaliate against a negative review within the review system. We study the effects of this treatment, first studied in the lab by Bolton, Greiner and Ockenfels (2012), on the types of reviews that are submitted and the subsequent demand for treated sellers.

The treatment had effects on the rate, speed, and types of reviews. The treatment increased guest reviews of hosts by 1.7% and host reviews of guests by 9.8% and reduced the time between checkout and review. We hypothesize that the treatment effects we observe are largely driven by a mechanism that has not previously been documented — users want to see and reveal what the other user wrote. In support of this hypothesis, we show that while the treatment decreased the amount of time before the first review is left by 9.7%, it decreased the amount of time between the first and second review by 35%. The change in review timing is seen for guests and hosts across all levels of experience. This is true even though we would expect that less experienced users have the highest incentive for reviews to be revealed since initial reviews are especially important (e.g. Pallais (2014), Farronato et al. (2019)). As a result, we argue that this effect may be due to curiosity in addition to the strategic incentive to have the review publicly visible on the platform.

The treatment also changed the types of reviews that were submitted. The ratings in the treatment were more negative on average but the effects were small — the average guest rating was just 0.25% lower in the treatment. We are also able to measure treatment effects on the review text. The incidence of negative review text increased by 12% for guests and 17% for hosts. However, since the baseline rates of negative text were 8.7% for guests and 1.9% for hosts, only a small share of transactions were induced to be negatively reviewed by the treatment.

Our experiment allows us to test a key prediction of prior work (Bolton, Greiner and Ockenfels (2012)) — that the simultaneous reveal treatment will reduce retaliation and will decrease the correlation between numerical ratings. We find evidence of a decrease in realized retaliation in the treatment. The rates of the worst (one star) ratings decreased by 31% for guests and this effect was concentrated on transactions where the host submitted negative text first. The treatment also decreased the correlation between review valence by 48% when measured with ratings and by 50% when measured with text.

We note that changes in reviews cannot solely be attributed to changes in strategic behavior. The reason is that the treatment increased review rates and consequently selection effects are likely to be present (Dellarocas and Wood (2007)). We use the methodology of principal stratification (Ding and Lu (2017)) to show that the treatment changed the reviewing behavior of individuals who would have reviewed regardless of the treatment (always takers). This, in addition to our results regarding the correlation of ratings, provides evidence that the effects on reviews are not caused solely by selection.

Lastly, we consider the effects of the treatment on subsequent outcomes on the platform. If the reputation system became more informative due to simultaneous reveal, then treated sellers (and especially those of lower quality) should see less demand or should invest more in quality. We do not detect causal effects of the treatment on subsequent host outcomes. We hypothesize that this lack of detectable effect is due to the small overall effect of the treatment on the realized distribution of ratings. We also look for heterogeneous effects across seller types and find no evidence that ex-ante worse hosts are hurt by the treatment. Our findings contrast with the results in Bolton, Greiner and Ockenfels (2012) and Hui, Saeedi and Sundaresan (2019) who find that similar changes to the reputation system decreased demand for low quality sellers. We discuss these papers in greater detail in section 2.

The rest of this paper proceeds as follows. In section 2 we describe the related literature in greater detail. Next, in section 3 we discuss several causal mechanisms that may operate in our setting and the predictions they generate for the effects of our treatment. Section 4 describes the setting of Airbnb and the design of the experiment. In section 5 we discuss the experimental design and in section 6, we discuss treatment effects and evidence regarding causal mechanisms. Section 7 contains the results of robustness checks and section 8 contains results pertaining to the effects of the experiment on adverse selection. Lastly, we conclude and discuss the implications of our results for reputation system design.

2 Literature Review

We contribute to three related research themes within the study of reputation systems. The first research theme studies why people submit feedback and whether this voluntary process produces bias. The second research theme concerns the effects of reputation system design on submitted reviews and subsequent market outcomes. The third research theme concerns reciprocity and trust on digital platforms including Airbnb.

Because the majority of reputation systems do not contain a payment scheme, the number, accuracy, and selection of realized reviews is determined by behavioral factors and the details of a particular reputation system. Avery, Resnick and Zeckhauser (1999) show that evaluations will be under-provided in equilibrium without an appropriate payment scheme and Miller, Resnick and Zeckhauser (2005) show how to design a scoring system with accurate reporting of feedback in equilibrium. These factors have been shown to matter in practice. Dellarocas and Wood (2007) argue, using data from eBay, that people with worse experiences are less likely to submit feedback. Subsequently, Nosko and Tadelis (2015), Cabral and Li (2014), Lafky (2014), Fradkin et al. (2015), and Brandes, Godes and Mayzlin (2019), have used experiments with rankings, coupons, and reminders to provide evidence for this hypothesis and the complementary hypothesis that people with more extreme experiences are more likely to review.

There are other reasons that the reviews collected by a reputation system may be biased. Li and Hitt (2008) argue that early buyers may have different preferences than late buyers, which could cause early reviews to be non-representative. Bondi (2019) provides a model and empirical evidence of this phenomenon in the market for books. Filippas, Horton and Golden (2018) argue that because reviewers may feel bad hurting a counterparty via a negative review, average review scores may inflate over time on platforms.

There have also been a number of studies focused on the effects of different reputation system designs. On Airbnb and similar markets, there is potential for adverse selection and moral hazard on both sides of the market. This fact makes it useful to have a two-sided reputation system. However, two-sided reputation systems may also allow for the conditioning of feedback on a counterparty's first rating, which can create biased feedback due to reciprocation and retaliation. Therefore, market designers may face a tradeoff between two-sidedness and bias. Three papers (Bolton, Greiner and Ockenfels (2012), Klein, Lambertz and Stahl (2016), and Hui, Saeedi and Sundaresan (2019)) study these tradeoffs.[1]

[1] Reciprocity has also been studied in other digital settings. Lorenz et al. (2011) use an experiment to show how adding social information to a wisdom of crowds task increases bias and Livan, Caccioli and Aste (2017) finds evidence of reciprocity in content platforms.

Bolton, Greiner and Ockenfels (2012) use data from several platforms as well as from laboratory experiments to document retaliation in two-sided review systems. They find that mutually negative feedback disproportionately occurs when sellers negatively review immediately after buyers negatively review, which is consistent with retaliation. The authors propose a simultaneous reveal system, like the one studied in our paper, and test it in the lab. They find that simultaneous reveal decreases ratings, reviews, and the correlation between ratings.

We conduct and analyze the first field experimental test of such a system. We find small effects on ratings, *increases* in the number of reviews, and decreases in the correlation of ratings. The differences in our results highlight important trade-offs between field and lab experiments. On the one hand, lab experiments may miss important features of the economic environment of a proposed policy (Levitt and List (2007)). On the other hand, as we discuss in section 7, the experiment we study is not as 'clean' as a laboratory experiment due to the practical considerations involved in running a large scale experiment with a company. Differences between the lab and our field setting include the social nature of the transaction, the underlying distribution of transaction quality, differences in how information was conveyed, the salience of notifications to review, and the incentive to have reviews revealed quickly. An important driver of our results is the desire to have review information revealed. This mechanism is not present in the laboratory experiments of Bolton, Greiner and Ockenfels (2012).

Klein, Lambertz and Stahl (2016) and Hui, Saeedi and Sundaresan (2019) study the effects of eBay's change from a two-sided to a (mostly) one-sided reputation system using a before and after observational study. We discuss these papers jointly since they are similar and provide important evidence on the effects of reputation system design. Klein, Lambertz and Stahl (2016) argue that the main effect of the change was to reduce strategic bias as measured by retaliatory feedback. They then argue that this reduction in bias leads to a decrease in moral hazard, as measured by an increase in realized post-transaction ratings. In contrast, Hui, Saeedi and Sundaresan (2019) argue that the improvement in measured buyer satisfaction is due to a reduction in adverse selection. Namely, after the change, low-quality sellers are less demanded even if they don't exit the market. Our paper complements these papers by studying a related policy in a different but equally important market. Furthermore, we use a randomized control trial which reduces concerns regarding the internal validity of the study. We do not find evidence that adverse selection was substantially reduced by this policy change.

We find that the distribution of ratings and the rates of reviewing changed due to the simultaneous reveal treatment. This calls for caution in using realized ratings to measure quality. In both Klein, Lambertz and Stahl (2016) and Hui, Saeedi and Sundaresan (2019), quality is primarily measured through changes in realized detailed seller ratings (DSRs). These papers argue that it is unlikely that the switch to a one-sided system affected DSR reviewing behavior, since DSRs are anonymous and only displayed as averages. Airbnb's star ratings are, like eBay DSRs, anonymous and only displayed as averages to hosts during our study period. We find that these star ratings are affected by the simultaneous reveal treatment, even for the first transaction in the experiment during which there is no possibility of a reduction in moral hazard or adverse selection. Therefore, for Airbnb and similar platforms, ratings cannot be used to measure changes in quality without an explicit model of reviewing behavior.

Lastly, reputation on Airbnb has been the subject of other academic studies. For example, Zervas, Proserpio and Byers (2015) compared ratings of the same accommodation on Airbnb and other digital platforms which have one sided reviews, and found that ratings on Airbnb are higher. We show that strategic considerations do not explain these differences since they have small effects on ratings. An earlier version of this paper, initially presented in 2014, contained many of the results presented in this work, and has influenced subsequent research regarding reputation on Airbnb including Proserpio, Xu and Zervas (2018) and Jaffe et al. (2017). Proserpio, Xu and Zervas (2018) propose that the social aspect of Airbnb transactions may affect realized quality in addition to reviewing behavior, while Jaffe et al. (2017) show how transactions with low quality sellers reduce guests' subsequent usage of the Airbnb platform.

3 Theoretical Framework

We now describe the factors that may drive differences in reviewing behavior between the baseline and simultaneous reveal (SR) reputation mechanisms. The simplest model of reviewing behavior is one in which guests and hosts submit reviews independently of each other's reviews. In this scenario, there would be no differences in observed reviews between the two reputation systems. Below, we discuss two deviations of reviewing behavior from the null model and their implications for the expected treatment effects of the simultaneous reveal policy. While both of these factors are likely to operate in our setting, the goal of our study is to learn which are quantitatively relevant.

The first potential mechanism is that guests and hosts reciprocate each other's reviews. That is, if a guest sees a positive host review, then they leave a more positive review than they would have left otherwise (and vice versa for negative reviews).

Implications of Reciprocity and Retaliation for the Effects of Simultaneous Reveal.

1. The correlation between the review ratings of the guest and host will decrease in SR (no information about first review).
2. Ratings and text will be lower on average in SR.
3. The incidence of one star ratings will decrease in SR due to reduced retaliation.
4. Review rates will decrease in SR because there is less incentive to review.
5. The time between reviews will increase in SR.

We now describe our reasoning for these predictions in detail. SR turns off reciprocity and retaliation in reviews. As a consequence, behavior caused by this mechanism should no longer be present in the treatment. We first consider the correlation between guest and host reviews. Since the second reviewer can no longer see the first review prior to writing their own review, the correlation of the valence (text and ratings) between the two reviews should fall.

The effect of reduced reciprocity and retaliation on the average rating depends on the relative rates of negative versus positive reciprocity. Since reviews tend to be overwhelmingly positive, we believe that the average effect of reciprocity on Airbnb is to increase ratings. Therefore, removing reciprocity should lower average ratings.

We should also be able to detect a decrease in retaliation in the treatment. Users who retaliate may want to do so in a way to maximally punish the counterparty. This is consistent with a grim-trigger strategy in repeated games. Cooperation is most likely to be sustained with the maximal possible threat of punishment. In the case of reviewing, cooperation occurs when both parties review each other positively. If the first reviewer does not review positively, the second reviewer can maximally punish them with a one star rating. Consequently, we expect that the treatment reduces the rates of one star second reviews since the second reviewer no longer knows if the first review is worth retaliating against.

Removing reciprocity also changes the incentive to review. Without seeing a positive review, the second reviewer has less of an obligation to reciprocate. This should reduce review rates and the speed of second reviews. Users may also strategically choose whether and when to review in order to induce reciprocity or avoid retaliation. This is no longer possible in the treatment. Recall that positive experiences (as measured by ratings) are an order of magnitude more frequent than negative experiences. Therefore, the incentive to induce reciprocity should overwhelm the fear of retaliation in terms of overall review rates. Indeed, Bolton, Greiner and Ockenfels (2012) find a reduction in review rates due to SR in their lab experiment. A similar line of reasoning implies that second reviews should occur more slowly without reciprocity.

The second (and novel to this paper) mechanism is that a user wants to see and show the other user's review. This may be driven by simple curiosity — a user wants to see what the other user said — or by the benefit of having a review revealed to other users on the platform. Because having more positive reviews increases demand, there is a business advantage in having reviews visible on the platform earlier rather than later. Below we list and explain the implications of this mechanism:

Implications of the Desire to Reveal the Review for the Effects of Simultaneous Reveal.

1. Review rates increase.
2. The time between reviews decreases. Furthermore, the hazard of second reviews after a first review should be higher in SR.
3. Ratings are lower on average.

Consider a user who has received a notification that the counterparty has reviewed. The user can immediately unlock this feedback by writing their own review, otherwise they must wait until the 14 day review period has ended. Since users prefer the information to be revealed, they now have an incentive to review and therefore review rates will increase. Additionally, users who see a first review in the treatment have more incentive to review quickly than users who see a first review in the control. This implies that the amount of time between first and second reviews should decrease and that the first review should have more of an effect on the hazard of reviewing by the second reviewer. The third prediction stems from the fact that people with worse experiences are generally less likely to review (as in Dellarocas and Wood (2007)). Therefore, additional reviews generated due to curiosity are likely to come from those individuals with worse experiences. Consequently, the marginal ratings should be lower on average.

4 Setting

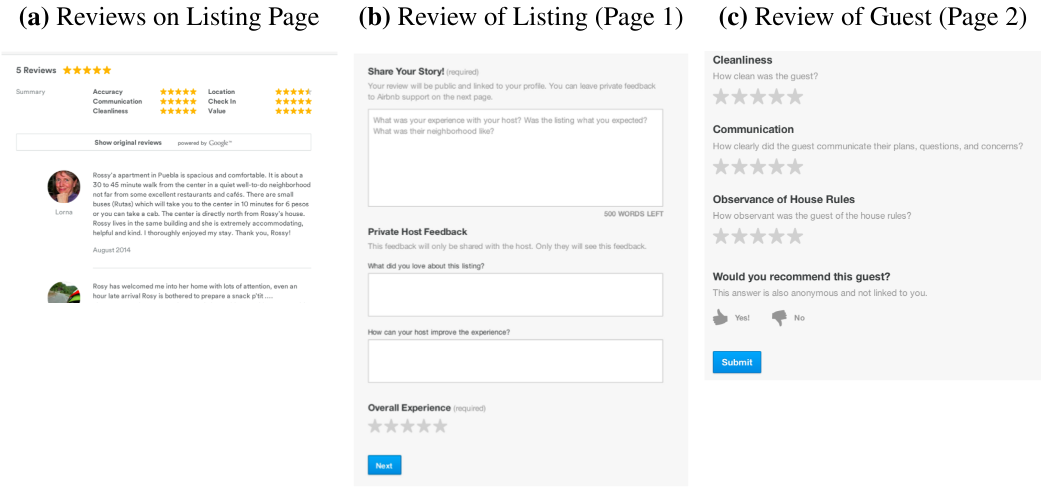
Airbnb describes itself as a trusted community marketplace for people to list, discover, and book unique accommodations around the world. Airbnb has intermediated over 400 million guest stays since 2008 and lists over five million accommodations. Airbnb has created a market for a previously rare transaction: the rental of an apartment or part of an apartment for a short term stay by a stranger.

In every transaction, there are two parties - the “Host”, to whom the listing belongs, and the “Guest”, who has booked the listing. After the guest checks out of a listing, there is a period of time (equal to 14 days for the experimental analysis and 30 days for the pre-experimental sample) during which both the guest and host can review each other.[2] Both the guest and host are prompted to review via e-mail the day after checkout. The host and guest are also shown reminders to review their transaction partner if they log onto the Airbnb website or open the Airbnb app (and may receive notifications related to reviews). Reminders are also sent when the counter-party submits a review, or if the reviewer has not left a review after certain, pre-determined lengths of time. Users cannot change their reviews after they have been submitted.

We now describe how reviews were solicited during 2014. Airbnb's prompt for guest reviews of listings consisted of two pages asking public, private, and anonymous questions (shown in Figure 1). On the first page, guests were asked to leave feedback consisting of publicly displayed text, a one to five star rating,[3] and private comments to the host.

The next page asked guests to rate the listing in six specific categories: accuracy of the listing compared to the guest's expectations, the communication of the host, the cleanliness of the listing, the location of the listing, the value of the listing, and the quality of the amenities provided by the listing. Rounded averages of the ratings were displayed on each listing's page once there were at least three submitted reviews. The second page also contained an anonymous question that asked whether the guest would recommend staying in the listing being reviewed. Overall ratings and review text were required and logged more than 99.9% of the time conditional on a guest review.[4]

Figure 1: Review flow on the website



(a) Reviews on Listing Page

5 Reviews ★★★★★

Summary	Accuracy	★★★★★	Location	★★★★★
	Communication	★★★★★	Check In	★★★★★
	Cleanliness	★★★★★	Value	★★★★★

Show original reviews powered by Google™

Lorna

Rossy'a apartment in Puebla is spacious and comfortable. It is about a 30 to 45 minute walk from the center in a quiet well-to-do neighborhood not far from some excellent restaurants and cafés. There are small buses (Rutas) which will take you to the center in 10 minutes for 6 pesos or you can take a cab. The center is directly north from Rossy's house. Rossy lives in the same building and she is extremely accommodating, helpful and kind. I thoroughly enjoyed my stay. Thank you, Rossy!

August 2014

Rosy has welcomed me into her home with lots of attention, even an hour late arrival Rosy is bothered to prepare a snack p'tit ....

(b) Review of Listing (Page 1)

Share Your Story! (required)

Your review will be public and linked to your profile. You can leave private feedback to Airbnb support on the next page.

What was your experience with your host? Was the listing what you expected? What was their neighborhood like?

500 WORDS LEFT

Private Host Feedback

This feedback will only be shared with the host. Only they will see this feedback.

What did you love about this listing?

How can your host improve the experience?

Overall Experience (required)

☆☆☆☆☆

Next

(c) Review of Guest (Page 2)

Cleanliness

How clean was the guest?

☆☆☆☆☆

Communication

How clearly did the guest communicate their plans, questions, and concerns?

☆☆☆☆☆

Observance of House Rules

How observant was the guest of the house rules?

☆☆☆☆☆

Would you recommend this guest?

This answer is also anonymous and not linked to you.

Yes! No

Submit

The review prompt for host reviews of guests was slightly different. Hosts were asked whether they would recommend the guest (yes/no), and to rate the guest in three specific categories: the

[2] There were some cases where a review was submitted after the 14 or 30 day time period. This occurred due to the manner in which emails were batched relative to the timezone, modifications to the trip parameters, or bugs in the review prompt system.

[3] In the mobile app, the stars are labeled (in ascending order) "terrible", "not great", "average", "great", and "fantastic". The stars are not labeled on the main website during most of the sample period.

[4] See Appendix A for additional details on the logging of review related data.

communication of the guest, the cleanliness of the guest, and how well the guest respected the house rules set forth by the host. Hosts were not asked to submit an overall star rating. The answers to these questions are not displayed anywhere on the website. Hosts also submitted written reviews that are publicly visible on the guest's profile page. Finally, the hosts could provide private text feedback about the quality of their hosting experience to the guest and separately to Airbnb.

5 The Simultaneous Reveal Experiment

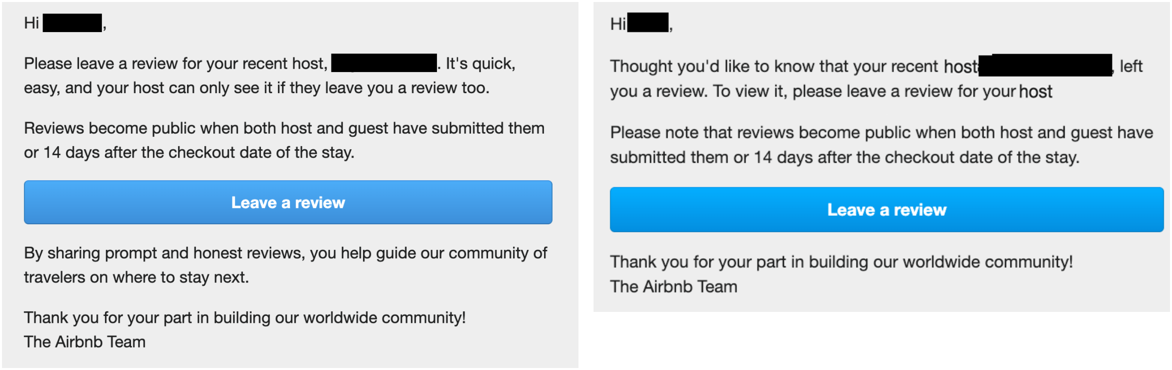
We now describe the design and effects of the simultaneous reveal experiment. Prior to May 8, 2014, both guests and hosts had 30 days after the checkout date to review each other and any submitted review was immediately posted to the website. This allowed for the possibility that the second reviewer retaliated against or reciprocated the first review. Furthermore, because of this possibility, first reviewers could strategically choose to not review or attempt to induce a reciprocal response by the second reviewer.

The experiment precluded this form of reciprocity by changing the timing with which reviews are publicly revealed on Airbnb. Starting on May 8, 2014, one third of hosts were assigned to a treatment in which reviews were hidden until either both guest and host submitted a review or 14 days had expired. Another third of hosts were assigned to a control group where reviews were revealed as soon as they were submitted and there was a 14 day review period.[5] Reviews were solicited via email and app within a day of the guest's checkout. An email was also sent when a counterparty submitted a review. Lastly, a reminder email was sent close to the end of the review period.

Users in the treatment received different reviews-related emails than users in the control. Figures 2 and 3 show the emails received by guests upon the end of their stay and when the counterparty left a review first. Figure 4 shows the analogous first email for hosts. During the simultaneous

[5] A final third were assigned to the status quo before the experiment, in which reviews were released as soon as they were submitted and there was a 30 day review period. We do not focus on the status quo in this paper because the difference in the reviewing period may have had an effect separate from the simultaneous reveal mechanism.

Figure 2: Simultaneous Review Email - Guest



(a) First Email

Hi [redacted],

Please leave a review for your recent host, [redacted]. It's quick, easy, and your host can only see it if they leave you a review too.

Reviews become public when both host and guest have submitted them or 14 days after the checkout date of the stay.

By sharing prompt and honest reviews, you help guide our community of travelers on where to stay next.

Thank you for your part in building our worldwide community!
The Airbnb Team

(b) Second Email

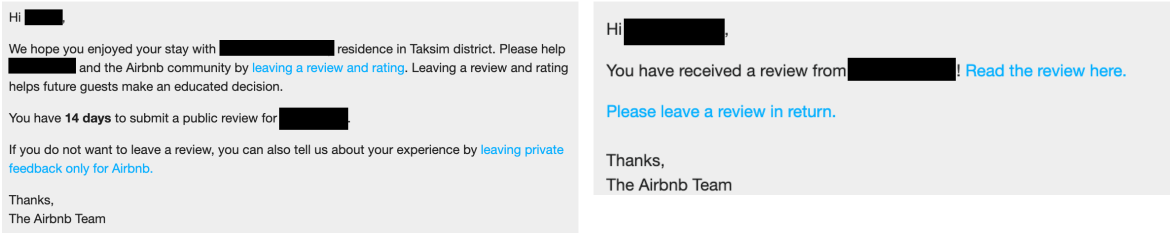
Hi [redacted],

Thought you'd like to know that your recent host [redacted], left you a review. To view it, please leave a review for your host

Please note that reviews become public when both host and guest have submitted them or 14 days after the checkout date of the stay.

Thank you for your part in building our worldwide community!
The Airbnb Team

Figure 3: Control Email - Guest



(a) First Email

Hi [redacted],

We hope you enjoyed your stay with [redacted] residence in Taksim district. Please help [redacted] and the Airbnb community by leaving a review and rating. Leaving a review and rating helps future guests make an educated decision.

You have **14 days** to submit a public review for [redacted].

If you do not want to leave a review, you can also tell us about your experience by leaving private feedback only for Airbnb.

Thanks,
The Airbnb Team

(b) Second Email

Hi [redacted],

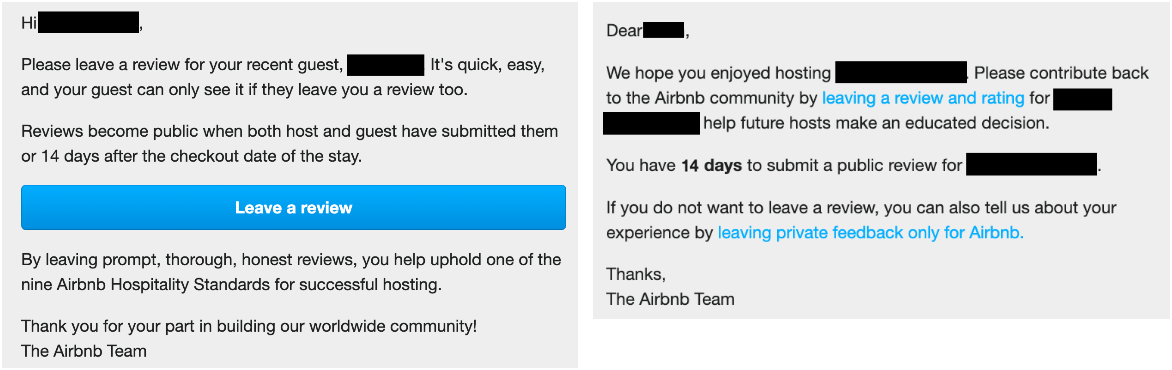
You have received a review from [redacted]! Read the review here.

Please leave a review in return.

Thanks,
The Airbnb Team

reveal experiment, Airbnb was also making unrelated changes to the content of reviews-related e-mails. In section 7, we discuss the potential impact of these changes on our results. Furthermore, both guests and hosts received a prominent notification before starting a review (Figure 5).

Figure 4: Host First Emails



(a) Treatment

Hi [redacted],

Please leave a review for your recent guest, [redacted] It's quick, easy, and your guest can only see it if they leave you a review too.

Reviews become public when both host and guest have submitted them or 14 days after the checkout date of the stay.

By leaving prompt, thorough, honest reviews, you help uphold one of the nine Airbnb Hospitality Standards for successful hosting.

Thank you for your part in building our worldwide community!
The Airbnb Team

(b) Control

Dear [redacted],

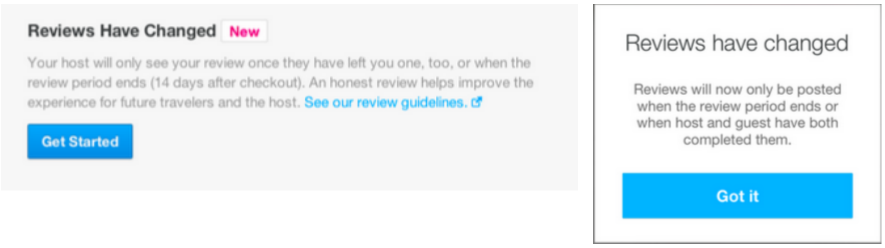
We hope you enjoyed hosting [redacted]. Please contribute back to the Airbnb community by leaving a review and rating for [redacted] help future hosts make an educated decision.

You have **14 days** to submit a public review for [redacted].

If you do not want to leave a review, you can also tell us about your experience by leaving private feedback only for Airbnb.

Thanks,
The Airbnb Team

Figure 5: Simultaneous Reveal Notification



(a) Desktop

Reviews Have Changed New

Your host will only see your review once they have left you one, too, or when the review period ends (14 days after checkout). An honest review helps improve the experience for future travelers and the host. See our review guidelines.

(b) Mobile

Reviews have changed

Reviews will now only be posted when the review period ends or when host and guest have both completed them.

The above figures display the notifications shown to guests prior to seeing the review form. For hosts, the desktop notification had the word ’host’ replaced with the word ’guest’.

5.1 Description of Reviewing Behavior in the Control Group

We now describe reviewing behavior in the 14 day control and the simultaneous reveal treatment. We focus on the first transaction observed for each host either in the treatment or in the control. We do this since we are, for the time being, interested in the effects of the experiment on reviewing behavior rather than on adverse selection or moral hazard, which may affect subsequent transactions. Since guests and hosts in this sample did not know about the change to the review system before the trip, they cannot adjust their matching or behavior during the trip to the new policy. In contrast, for subsequent transactions, the treatment may affect selection and effort during the transaction. Furthermore, this sample restriction allows us to avoid issues due to spillovers between multiple listings managed by the same host. We later turn to the effects of the experiment on subsequent stays and reviews.

Table 1: Summary Statistics

	Guest			**Host**		
	Control Mean	*Treatment Mean*	*Effect*	*Control Mean*	*Treatment Mean*	*Effect*
Submits Review	0.68	0.69	0.01 ***	0.72	0.79	0.07 ***
Recommends	0.97	0.97	0.00	0.99	0.99	0.00 *
Overall Rating = 5	0.74	0.73	-0.01 ***			
Lowest Subrating = 5	0.48	0.47	-0.01 ***	0.82	0.81	-0.01 ***
Review Text Negative	0.13	0.14	0.01 ***	0.03	0.03	0.00 **
Lowest Subrating	4.29	4.27	-0.02 ***	4.79	4.78	-0.02 ***
Time to Review (Days)	4.70	3.89	-0.81 ***	3.80	3.42	-0.37 ***

	Transaction		
	Control Mean	*Treatment Mean*	*Effect*
Days Between Reviews	3.05	1.98	-1.07 ***
Both Review	0.55	0.62	0.07 ***

This table displays mean outcomes in the control and treatment, as well as treatment effects, at the guest, host, and transaction level. The rating related outcomes are computed conditional on a review. The effect is displayed in percentage points. *p<0.1;**p<0.05;***p<.01;

Our baseline sample consists of 119,789 transactions starting with checkout dates on May 10, 2014 and ending with checkout dates on June 12, 2014.[6] Table 1 displays the mean outcomes for the control and treatment groups.

In the control group 68% of trips result in a guest review and 72% result in a host review. Reviews are typically submitted within a few days of the checkout, with hosts taking an average of 3.8 days to leave a review and guests taking an average of 4.7 days. Reviews are mostly positive. Conditional on a review, 74% of guests leave a five star overall rating and 48% of guests submit fives for all of the category ratings. Figure 6a displays the distribution of ratings for reviews by guests (red) and hosts (blue). Both distributions are skewed towards the right, with the majority of ratings being four and five stars. Host reviews are even more positive than guest reviews, with 82% of host reviews containing all five star category ratings. Both guests and hosts submit positive recommendations over 97% of the time, conditional on an observed submission of a recommendation. These high recommendation rates are notable, given that the answers are anonymous and are never seen by anyone other than Airbnb.

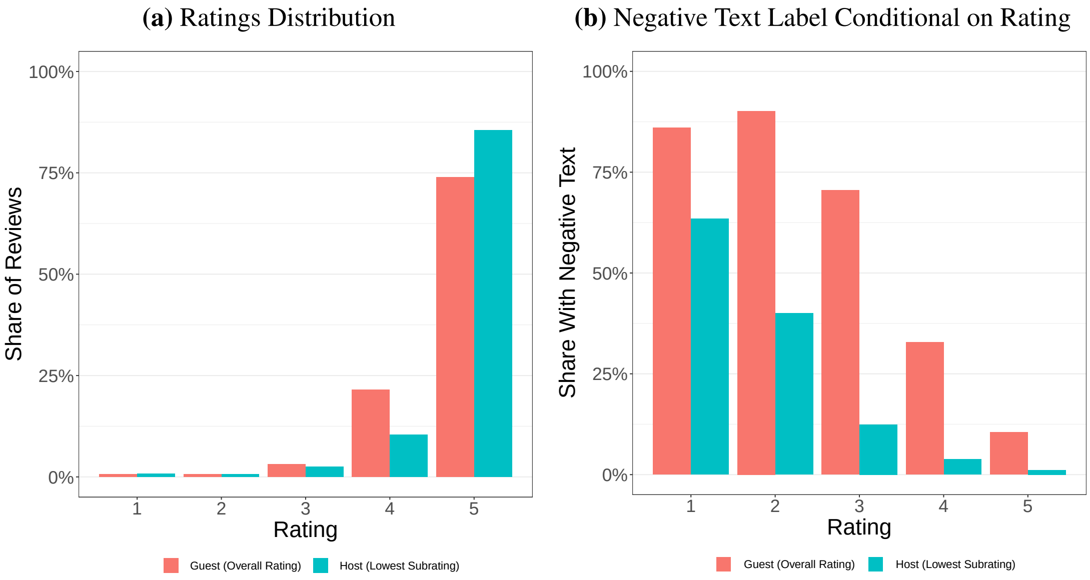
Text comprises another important part of the review which we incorporate into our analysis. We trained a regularized logistic regression model on pre-experiment data to classify the sentiment of reviews and to determine the words and phrases associated with negative reviews. A discussion of the training procedure can be found in Appendix B.

In Figure 6b we show the share of negatively labeled text reviews by star rating in the control group. Low star ratings by guests are typically but not always associated with negative text. Over 89% of one and two star reviews by guests are classified as negative while three star reviews have text that is classified as negative over 70% of the time. Hosts are less willing to leave negative text even when they leave a low category rating for the guest.

With regards to more positive reviews, negative text is less prevalent but still exists. Guests write negatively classified text 32% of the time for four star reviews and 9.8% of the time for

[6] Although randomization began for trips ending on May 7, 2014, we exclude trips with checkouts between May 7, 2014 and May 9, 2014 due to inconsistencies in logging treatment assignments on those days. Appendix A recreates our main results with a sample that excludes any host with a trip ending on these days. This appendix also includes details regarding treatment assignment logging issues on June 6, 2014 and June 7, 2014. Because we only analyze each host's first trip during the experiment and this span of days occurs toward the end of the experiment, these logging issues do not substantively affect our results. Note that the experiment ran all the way until the public announcement and launch of the policy to the entire platform. We do not use data from close to the launch in our main analysis because reviewing behavior may have been affected by the launch.

Figure 6: Ratings Distributions



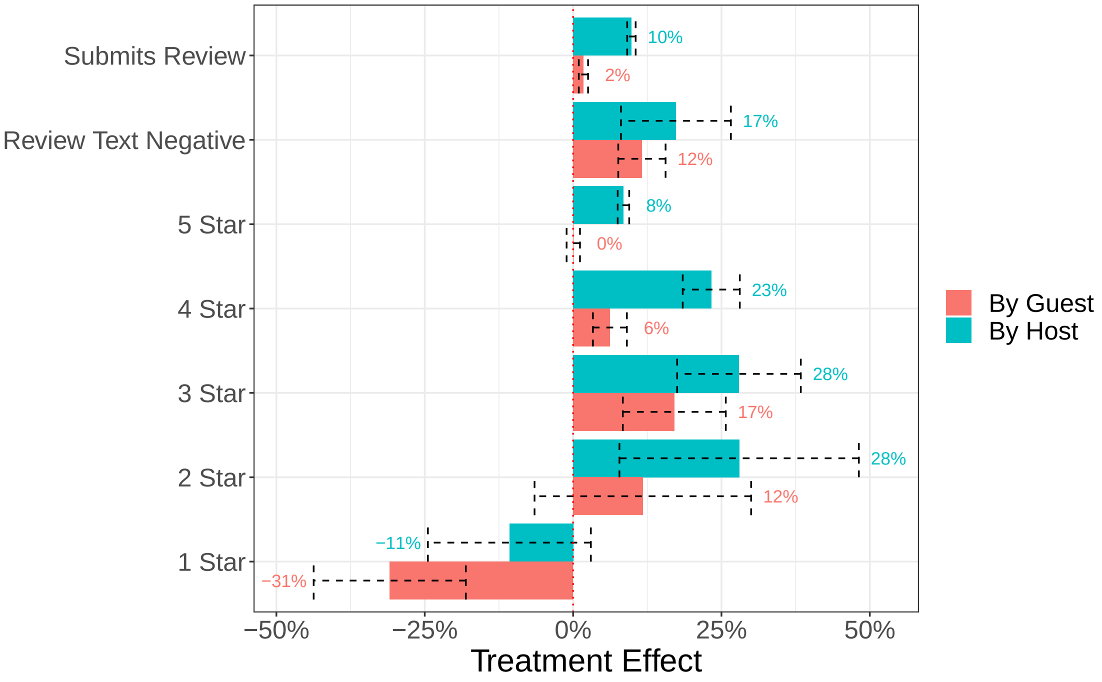
The left figure displays the distribution of submitted overall ratings by guests and lowest category ratings by hosts in the control group of the simultaneous reveal experiment. The right figure displays the prevalence of negative text review as predicted by a regularized logistic regression conditional on rating.

five star reviews. This may be due to the desire for guests to explain shortcomings, even if they had a good experience. Another explanation, especially relevant to five star reviews, is that text classification does have some measurement error.

6 The Effects of the Experiment

We now present the main effects of the experiment and discuss the mechanisms that drive these effects. We begin by analyzing the treatment effect on guest reviews of hosts (Figure 7). The experiment increased the number of guest reviews by 1.7%. The effects were non-monotonic in rating types. There was no detectable effect on the number of five star ratings and at least a 6.2% increase in the number of two, three, and four star ratings (although the increase in two star ratings is not statistically significant). The number of one star reviews fell by 31% in the treated group. Lastly, the rate of reviews with negative text increased by 12%.

Figure 7: Effects of Experiment on Reviews



This figure displays the percentage change (relative to the mean in the control group) in reviews of a given type due to the treatment. The standard errors used for the 95% confidence intervals are calculated using the delta method. Transactions with no label, such as when there is no review, are treated as zeros for the purpose of this calculation.

Figure 7 also displays analogous results for hosts.[7] The review rate went up by 10% and the rate of reviews with negative text went up by 17%. Similar to the pattern seen with guests, the largest relative increase came from reviews with two through four stars. The treatment also increased the number of five-star ratings by 8.5%, and decreased the number of one-star ratings by 11% (although this is not statistically significant). To summarize, the treatment increased review rates, caused more negative reviews on average, and decreased the number of one star reviews.

6.1 Reciprocity and Retaliation

In this section, we test our theoretical predictions regarding the reciprocity and retaliation mechanism. Recall that our first prediction in section 3 was that the treatment would reduce the cor-

[7] Numerical score effects are calculated with respect to lowest category ratings for hosts, since they do not submit an overall rating.

relation between guest and host review valence. We confirm this prediction by measuring the correlation in reviews in two ways: using the labeled text review and the lowest rating (including sub-ratings) for cases where we observe these measures for both reviews. Across both measures, we find large and statistically significant decreases in the correlation of ratings. The correlation of positive text fell by 50% (Std. Err.: 6.7%), and the correlation of ratings fell by 48% (Std. Err.: 4.4%).

We also confirm the second prediction of the theory – that the ratings and text will be lower on average. The third prediction is a fall in one star ratings, particularly for second reviews following negative reviews. Figure 7 documents a fall in one star ratings in the treatment. We also find a fall in one star ratings when the first review contained negative text. Specifically, one star ratings by guests following a negative host review decreased in the treatment by 66%. This happens even though we have 14% more observations with negative host reviews first and any guest review second in the treatment. We find similar results for host reviews after a negative guest review. The rate of host reviews with a one star rating fell by 65% in the treatment when guests left a negative textual review. Therefore, we find evidence for a grim-trigger strategy of retaliation in the control group.

However, predictions four and five of the reciprocity theory are not borne out in the experiment. We have already shown that review rates actually increase, contradicting prediction four. The timing of reviews is also inconsistent with a pure reduction in reciprocity and retaliation. The first review of a transaction is submitted 9.7% (Std. Err.: 0.57%) faster in the treatment, and faster reviews occur for both guests and hosts. Furthermore, theories of reciprocity predict that second reviews will come soon after first reviews. Since the treatment reduced reciprocity, we would expect that the time between guest and host reviews would increase in the treatment. In contrast, we find that the time between guest and host reviews decreases by 35% (Std. Err.: 0.72%).

6.2 The Desire to Reveal a Review

The hypotheses set forth in Section 3 imply that if the treatment only affected the amount of reciprocity on Airbnb, review rates would decrease and the time between reviews would increase. However, we observe treatment effects with the opposite signs in the data, suggesting that our experiment also interacts with other behaviors. One such behavior is the desire to reveal review information to oneself and/or to other users. Recall that reviews in the treatment are hidden until both parties review or until 14 days have elapsed. This means that users cannot immediately see a counterparty's review.

Users may want to reveal a counterparty's review for one of two reasons. The first of these reasons is that they are curious about the review. The second is that it is advantageous to have a review visible on the platform as soon as possible. Given that most reviews are positive and that positive reviews increase demand on the platform, hosts should be especially eager to have reviews revealed. If users knew that the first review triggered a quick second review, they would want to write their own first review quickly.

The time to review is naturally modeled as a duration. We would like to understand how the treatment affects the hazard of reviewing and whether the hazard is differentially affected by the counterparty's review in the treatment. Table 2 displays the estimates of Cox-proportional hazard models of review times. We find that the treatment increases the overall hazard of reviews for guests by 8%. Next, we interact the treatment with whether the host has already reviewed. We find that the hazard increases by 147% after a host review, and it does so an additional 12% in the treatment. Therefore, the treatment increases the speed with which guests review after a host reviews but the effect is modest compared to the baseline.

Table 2: Speed of Review Effects

	Dependent variable:				
	Relative Hazard of Review				
	Guest		Host		
	(1)	(2)	(3)	(4)	(5)
Treatment	1.080 t = 11.003***	1.014 t = 1.450	1.230 t = 31.003***	1.108 t = 12.560***	1.114 t = 6.489***
After Host		2.470 t = 86.862***			
Treat * After Host		1.123 t = 8.276***			
After Guest				1.989 t = 63.448***	2.137 t = 35.320***
Treat * After Guest				1.554 t = 30.948***	1.465 t = 13.067***
Treat * 1 - 3 Reviews					1.096 t = 2.230**
Treat * 4 - 12 Reviews					1.075 t = 1.794*
Treat * > 13 Reviews					1.073 t = 1.722*
Number of Events	82055	82055	90034	90034	90034
R^2	0.001	0.084	0.006	0.089	0.091

Note: *p<0.1; **p<0.05; ***p<0.01

This table displays the relative hazard estimated from cox proportional hazard regressions where the outcome is whether a review a submitted by the guest (columns 1 - 2) or the host (columns 3 - 5). ‘After Host’ and ‘After Guest’ refer to an indicator for whether the time was after the submission of the review by the counterparty. ‘Reviews’ in column (5) refer to the number of reviews for the listing at the time of the booking.

We conduct a similar exercise for host reviews of guests. We find that the treatment increased the hazard of host reviews by 23%. We also find that guest reviews increase the hazard of host reviews by 99% in the control and by an additional 55% in the treatment. The large increase in host review hazard after a guest review is consistent with the relatively large increase in overall review rates by hosts in the treatment.

Lastly, we consider whether the increase in reviews is caused by curiosity or by the strategic desire to reveal the review on the platform. First, to the extent that guests typically don't need many reviews in order to transact (Cui, Li and Zhang (2019)), they should be driven mostly by curiosity. Second, hosts with fewer reviews should have more incentive to have reviews revealed quicker. This is because the value of a review decreases with the number of prior reviews. Our prior results point to larger effects on host reviews, which favors the strategic explanation.

Column (5) displays the results from a full interaction between treatment, whether a guest has already reviewed, and the number of prior ratings a listing has received. We are not able to detect much heterogeneity, except that the effects of a counterparty review are smallest for listings with 0 prior reviews. We find similar results in a linear model where the outcome variable is whether a review by a user comes within a day after a review by the counterparty (Table AII).

The results for guests and hosts don't paint a clear picture regarding the motivations of users to review quickly after an initial review. On the one hand, the effects are generally bigger on host reviews, which suggests that the desire to publicly display reviews is important. On the other hand, hosts with more reviews are more influenced by guest reviews to quickly submit a second review even though they have less of an incentive for speeding up the review. The larger effect for more reviewed hosts suggests that the desire to public display reviews in order to improve market outcomes is not driving host behavior. We suspect that the rate of reviewing is likely to increase for both reasons - curiosity and desire for the information to be publicly revealed - but are uncertain of their relative importance.

6.3 Selection into Reviewing by Treatment Status

The effect that we observe on review ratings may be due to a change in the types of people who review (selection) or due to a change in the reviews users submit conditional on choosing to leave a review. Namely, if the typical experience of those who review in the treatment is different than the typical experience of those who review in the control, then this may explain differences in reviews across conditions. We've already shown that reviews change in ways consistent with a reduction in reciprocity and retaliation, which is not solely driven by selection. Below, we provide additional evidence that the effects we find are due to changes in reviewing behavior rather than just changes in who reviews.

The first piece of evidence showing that the effects we find are not caused by selection is that the number of reviews induced by the treatment is relatively small. For example, there are only 2% more reviews by guests in the treatment. One would need an extreme level of selection for this 2% to account for all of the changes in the ratings that we observe.

To be more precise, suppose that we assume monotonic effects of the treatment on review rates. That is, we posit that the treatment weakly increases review rates for all individuals. This is consistent with the increases in review rates we see for both guests and hosts. Conditional on this assumption, decreases in the number of one star ratings cannot be caused by selection.[8] Selection would also not be able to explain cases when the increase in a review type exceeds the total increase in reviews. For example, the number of guest reviews increases by 1.1pp while the number of guest 3 and 4 star reviews increases by 1.2pp. Under monotonicity, it must be the case that at least some individuals changed the types of reviews which they submit due to the treatment.

We also test for changes in reviewing behavior in a more statistically rigorous way using the principal stratification framework (Frangakis and Rubin (2002) and Ding and Lu (2017)), which allows for the estimation of the causal effect of a treatment for two different subpopulations: always takers (those who review regardless of treatment status) and compliers (those who review only if treated). Any effects on the always-takers are not driven by selection. The key assumption

[8] Note that the rate of one star reviews per transaction fell in the treatment.

required for implementing principal stratification is weak general principal ignorability (Ding and Lu (2017). It states that the expected outcome, conditional on submitting a review, is independent of strata (complier and always taker) when controlling for covariates.[9] This is a strong condition but is made more plausible by the availability of pre-treatment covariates such as historical ratings by guests and hosts, as well as trip characteristics including whether there were customer service complaints. We discuss the details of our principal stratification procedure in Appendix E.

We find that the treatment does change reviewing behavior for the always-takers — those individuals who would review regardless of treatment status. Figure 8 displays the causal effects for this set of users. We see a pattern of treatment effects consistent with our baseline results. The always takers are caused to submit more two - four star ratings relative to one or five star ratings as a result of the treatment and to leave more negative text. In other words, the treatment not only changed which Airbnb users left reviews, but how they reviewed their counterparty conditional on leaving feedback.

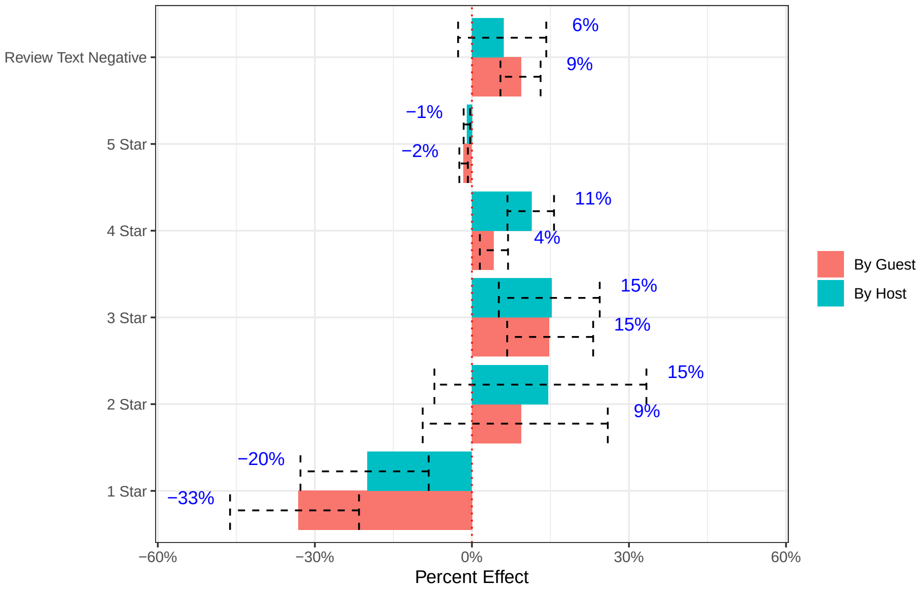
We are also able to estimate the distribution of ratings by compliers — those induced to review by the treatment. We measure the reviewing behavior of this group relative to the always taker group in Figure A5. We find that compliers tend to leave lower ratings relative to always takers. This is consistent with the work of Dellarocas and Wood (2007) and others who find that those who don't review tend to have worse experiences than those who do.

6.4 Learning About the Treatment

A final concern is that it may have taken some time for users to learn about the change in the review policy. For example, users may not have noticed that reviews had changed or that the change in reviews allowed them to be more honest when reviewing. First, note that the reviewers were presented with a prompt prior to the review which prominently explained the treatment and required a click after reading (Figure 5). Second, since our experiment was only conducted for

[9] An analogous assumption regarding never takers and compliers in the control holds trivially since they don't submit reviews.

Figure 8: Always Taker Causal Effects



This figure displays the percentage change (relative to the mean in the control) in reviews of a given type due to the treatment. The standard errors used for the 95% confidence intervals are calculated using a the percentile bootstrap method. Transactions with no label, such as when there is no review, are treated as zeros for the purpose of this calculation.

a short period of time, we may not have been able to capture the full extent of this learning. In Appendix F we document how reviewing behavior evolved following the launch of simultaneous reveal to the entire site in July of 2014. We find that the differences in reviewing behavior following the launch are consistent with our observed treatment effects, suggesting that learning about the policy over time is not first order for this setting.

7 Alternative Explanations of Experimental Effects

We now discuss threats to our interpretation of the experimental results. In our prior analysis we explained the simultaneous reveal treatment effects as a bundle of mechanisms including removing reciprocity or retaliation and introducing an incentive to review in order to reveal information. However, Airbnb's implementation of the treatment may have inadvertently introduced another reason for changes in the number and types of reviews. Recall that the emails in Figure 2 and Figure 3 differed not only in the information that they conveyed but also in the size of the 'Leave a review' button and the specific text in the email (i.e. 'Thank you for your part in building our worldwide community!'). It could be the case that these factors contributed to the effects which we observe.

7.1 Larger Button

Suppose that we were interested in studying a treatment where the only difference was the increased size of the 'leave a review' button in the treatment emails. Such a treatment would be likely to increase review rates. It would also likely decrease the average ratings since marginal reviewers are likely to have less extreme experiences than the typical reviewer (e.g. Dellarocas and Wood (2007), Fradkin et al. (2015), and Brandes, Godes and Mayzlin (2019)). Both of these effects are consistent with our treatment effects.

However, the 'bigger button' treatment is unlikely to explain our other results. First, note that users are prompted to review not only via email but also via app and whenever they log into the site. Since hosts are more likely to visit the site and the app due to their managing of bookings and stays for multiple guests, they are more likely to see non-email-based review prompts than guests. Therefore, any individual prompt is less likely to be pivotal for a host than a guest. In contrast, we find much larger effects on the review rates of hosts than guests.

Second, we find that the effect of an email about a counterparty's review is much bigger in the treatment than in the control when sent to hosts. In contrast, we do not see such pronounced differences in the effects of the first email. This cannot be explained by the bigger button, since the bigger button also appears in the first email prompt sent to the host.

Third, we argued in the previous section that many of the effects we observe are not solely driven by selection into reviewing. Since the primary effect of a 'bigger button' would be selection, the 'bigger button' cannot explain them.

7.2 E-mail Text

A related concern is that the exact email copy sent to users varied over the course of the experiment. To investigate this, we solicited Airbnb review emails, via social media, from the time period in question. For guests in the control group, we found three versions of the email, which varied in the color scheme and logo.[10] We believe that the difference in logo is due to a dynamic link in the email and that the users saw the old logo when they actually received the email during the experiment period. If this change were material, then we would expect to see differences in the treatment effect over time when the email copy changed. We find that the effect of the treatment on guest review rates does not substantially change over the course of the experiment (Figure A6).

Similarly, we found that Airbnb inserted an additional piece of content in some of the initial treatment emails sent to hosts (the exact time at which this began is unclear to us). This content describes how reviews have changed (Figure A3) and was deployed for some members of the treatment group.

We have several comments regarding this interstitial. First, to the extent that this interstitial was shown to some treated users and not others, our treatment effects reflect a mix of the effects of the two emails. Second, hosts were exposed to the change in policy in other ways, including app notifications, a prompt in the review flow, and calls to action on the website. These were not different across treated users. Furthermore, the reminder emails and emails about guest reviews were, based on our data collection efforts, constant across treated users during the treatment period. Lastly, if the interstitial was changed at a particular period in time during the experiment, then we can look for heterogeneous effects over time. We computed daily treatment effects on review rates and found that the effect is similar throughout the duration of the experiment (Figure A7).

[10] Airbnb introduced a new logo on July 16, 2014, which is after our experiment concluded.

8 Effects on Adverse Selection

We now discuss the effects of the treatment on the selection of transacting users. If the treatment had its intended effect, then transactions with low quality users should become less likely in the treatment and transactions with high quality users should become more likely (Airbnb (2014)). Prior observational and lab work about bilateral reputation systems has argued that removing retaliation and reciprocity reduces adverse selection for sellers (Hui, Saeedi and Sundaresan (2019)). We use our experiment to study the effects of simultaneous reveal on adverse selection in Airbnb and find precisely estimated null effects.[11]

We begin by describing the mechanisms by which simultaneous reveal may affect adverse selection. First, simultaneous reveal reviews were less influenced by reciprocity, which should in theory make them more reflective of user experiences. This more accurate information should create better (although possibly fewer) matches as it redistributes demand from worse listings to better listings.[12] However, the simultaneous reveal policy does not just cause an increase in review accuracy — it also increases the speed and total number of reviews due to the desire to reveal information. Since induced reviews are typically positive, this may cause an increase in demand for the treated listings which are more likely to be rated.

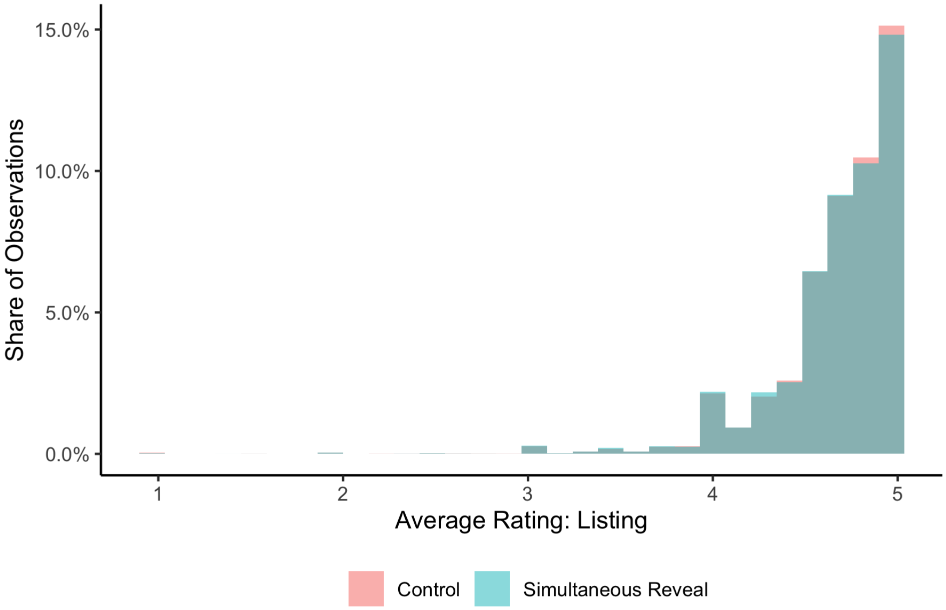
One way to measure the potential impact of the policy is to consider the distribution of average ratings for listings in the treatment and control groups after the first review. Because hosts have already accumulated many reviews, the initial effect of the policy on average ratings at a listing level is small. We plot the difference in realized average ratings for the treatment and control groups in Figure 9. We find small differences, with the control group having slightly more listings with an average rating closer to five stars.

Next, we measure the effects of the treatment on listing outcomes. We focus on two kinds of outcomes. The first are measured exclusively in the experimental period. During the experimental

[11] We can also reject large effects of the treatment on subsequent guest outcomes (Table AIII).

[12] Klein, Lambertz and Stahl (2016) propose a toy model of reviewing, retaliation, and market outcomes. In their model, eliminating retaliation induces more honest (lower) ratings and causes seller exit and increased effort provision. Because our treatment had effects on review quantity and speed in addition to reducing average ratings, these simple predictions do not necessarily apply to our setting.

Figure 9: Distribution of Average Ratings at a Listing Level



This figure displays the distribution of the average rating across reviews within a listing following the first transaction in the experiment.

period, the differences between reviews in the treatment groups should be most pronounced. However, there will have been less time for those reviews to affect subsequent guest and host behavior.

Table 3 shows precisely estimated zeros for the log of nights in the experimental period and log of average booked price per night in the experimental period. The estimates for the log of revenue are less precise but are still not statistically distinguishable from zero.

We then look at outcomes through the end of 2014. Note that since the treatment was launched platform-wide in July, 2014, both treatment groups were partially treated using this outcome metric. We find precisely estimated zeros on the log of bookings through 2015 and whether the listing is active in 2015. In summary, the exposure of listings to the simultaneous reveal treatment does not affect aggregate demand.

As discussed above, we also predict that worse quality listings should receive less demand than high quality listings as a result of the treatment.[13] Such a decrease in demand would represent a reduction in adverse selection. We propose two proxies for listing quality that are unaffected by the treatment and use these to test for heterogeneous treatment effects. Adverse selection would be reduced if ex-ante worse listings received less demand due to the treatment.

Table 4 displays the specifications that interact the treatment with measures of listing quality. We add two interaction variables. The first of these is the ratio of five star ratings to total transactions occurring prior to the experiment. We call this the effective positive percentage (EPP) as in Nosko and Tadelis (2015), who argue that this is a good proxy for quality. We also add an indicator for whether we can measure the EPP since it is undefined when there are no prior transactions. As intended, higher EPP is associated with better subsequent listing outcomes even in the control, meaning that it is a good proxy of listing quality. However, the interaction of this variable with the treatment is close to zero and not statistically significant. We conduct a similar exercise with another proxy of quality — the occurrence of a customer service complaint during the first transaction in the experiment.[14] We find that customer service complaints are associated with worse subsequent listing outcomes even in the control. However, we find no statistically significant interaction effects with the treatment. To summarize, we don't find decreases in adverse selection using two proxies for listing quality.

Table 3: Treatment Effects on Listing Outcomes

	Dependent variable:				
	Log(Nights in Exp.)	Log(Price in Exp.)	Log(Rev. in Exp.)	Log(Bookings by 2015)	Active in 2015
	(1)	(2)	(3)	(4)	(5)
Treatment	−0.010	−0.005	−0.025	0.002	−0.003
	(0.006)	(0.005)	(0.019)	(0.004)	(0.002)
Controls Included	Yes	Yes	Yes	Yes	Yes
Observations	119,550	73,234	119,550	119,550	119,550
R^2	0.262	0.411	0.219	0.631	0.078
Note:					*p<0.1; **p<0.05; ***p<0.01

This table displays the treatment effects on listing outcomes after the first transaction in the experiment. Controls are included for greater than median effective positive percentage (EPP), whether the EPP is calculable, log of prior bookings, log of the first price, and whether the guest submitted a customer service complaint. Columns with 'In Exp' in the name refer to outcome calculated only through June 12, 2014, the end of the experimental period. There are fewer observations for the price variable, because we can't measure transaction prices for hosts who did not transact after the initial transaction in the experiment.

[13] Another possibility is that the treatment reduces moral hazard, which we were unable to test for since we cannot measure quality. Using realized ratings, in the treatment, as measures of quality, as in the prior literature, is problematic since the treatment affects ratings in ways other than through quality.

[14] We exclude customer service complaints which occur after the transaction has finished since they may be affected by the treatment.

Table 4: Tests of Adverse Selection

	Dependent variable:				
	Log(Nights in Exp.)	Log(Price in Exp.)	Log(Rev. in Exp.)	Log(Bookings by 2015)	Active in 2015
	(1)	(2)	(3)	(4)	(5)
Treatment	−0.009	−0.007	−0.023	0.005	−0.001
	(0.009)	(0.007)	(0.029)	(0.006)	(0.004)
> Median EPP	0.067***	0.049***	0.222***	0.054***	0.039***
	(0.010)	(0.007)	(0.030)	(0.006)	(0.004)
No EPP	−0.034***	−0.047***	−0.282***	−0.023***	0.001
	(0.013)	(0.012)	(0.040)	(0.008)	(0.005)
Customer Service	−0.160***	−0.016	−0.484***	−0.108***	−0.036**
	(0.038)	(0.032)	(0.119)	(0.024)	(0.016)
Log(Num Prior Bookings)	0.335***	0.034***	0.869***	0.542***	0.038***
	(0.003)	(0.002)	(0.008)	(0.002)	(0.001)
Log(First Price)	−0.140***	0.339***	−0.298***	−0.140***	0.014***
	(0.003)	(0.003)	(0.010)	(0.002)	(0.001)
Treat * > Median EPP	−0.005	0.0005	−0.005	−0.008	−0.005
	(0.013)	(0.010)	(0.042)	(0.008)	(0.006)
Treat * No EPP	0.004	0.012	−0.007	0.002	0.002
	(0.016)	(0.015)	(0.051)	(0.010)	(0.007)
Treat * Customer Service	0.034	0.039	0.076	−0.043	−0.031
	(0.054)	(0.046)	(0.170)	(0.034)	(0.022)
Observations	119,550	73,234	119,550	119,550	119,550
R^2	0.262	0.411	0.219	0.631	0.078
Note:					*p<0.1; **p<0.05; ***p<0.01

This table displays the treatment effects on listing outcomes after the first transaction in the experiment. Controls are included for greater than median effective positive percentage (EPP), whether the EPP is calculable, log of prior bookings, log of the first price, and whether the guest submitted a customer service complaint. Columns with 'In Exp' in the name refer to outcome calculated only through June 12, 2014, the end of the experimental period. There are fewer observations for the price variable. This is due to the fact that we can't measure transaction prices for hosts who did not transact after the initial transaction in the experiment.

One may also be concerned that the small average treatment effects mask other types of heterogeneity not identified by our proxies for low quality listings. For example, a marginal positive or negative review my have large effects for a subset of listings. We test for this by comparing the distribution of bookings through 2014 between the treatment and control. Figure A8 shows that these distributions are very similar. We also conduct a Kolmogorov-Smirnoff test on the equality of these distributions and fail to reject the null (p-val = 0.6296). This confirms that earlier exposure to the treatment had, at most, negligible effects on average market outcomes. We discuss the implications of these findings in the next section.

9 Discussion

Reputation systems are an important component of a well-functioning online marketplace. However, because informative reviews are public goods, reputation systems don't capture all relevant information and observed ratings may be biased. These systems may be especially difficult to design for peer-to-peer markets in which services are exchanged. In these settings, market participants can review each other and may meet in person, resulting in reciprocity and retaliation within the review system. In this paper, we study the effects of a simultaneous reveal policy intended to reduce reciprocity and to improve market outcomes.

We find that the simultaneous reveal policy increased review rates and decreased the average valence of reviews. It also reduced retaliatory one star reviews as well as the correlation between guest and review ratings. The effects we find are due to at least two mechanisms - a reduction in reciprocity and a desire to reveal information. We also note that while the relative effects of the treatment on reviews are substantial — the treatment increased reviews with negative text by over 12% for both guests and hosts — the absolute effects are small. For example, negative review text by guests occurs in just 8.7% of transactions in the control, so only a small share of transactions are affected by this treatment.

The ultimate goal of reputation system changes should be to improve the quality of transactions in the market. For example, the intention of the simultaneous reveal policy was to make reviews more commensurate with experienced transaction quality, with the idea that more informative reviews will lead to better matches. Of the mechanisms we document, the reduction in reciprocity should indeed have this intended effect. On the other hand, the informative value of additional reviews induced by the desire to reveal information is uncertain. We study whether simultaneous reveal led to better matches and reduced adverse selection, and find that it did not. Note that the null effects we find may be driven by the fact that the experiment ran for a relatively short time prior to simultaneous reveal reviews being launched across the entire site.

We draw several lessons about reputation systems from our results. First, while it is widely known that review information can be biased, it is less acknowledged that magnitude of this bias can change over time due to changes in the reputation system design. This can be true even for aspects of the review that are anonymous and/or private and, consequently, expected to be less subject to bias. The simultaneous reveal treatment only affected the timing of the disclosure of review text to a counterparty. Nonetheless, the treatment changed both review text and star ratings.

Another lesson we draw is that real world reviewing behavior may be hard to replicate in a laboratory setting. The laboratory tests of the simultaneous reveal policy conducted by Bolton, Greiner and Ockenfels (2012) showed decreases in review rates whereas we found increases. We hypothesize that this is caused by the desire to reveal information, a motivation for reviewing not present in the laboratory experiment. Other potentially important differences between our setting and the lab include differences in the underlying distribution of transaction quality and the presence of social, rather than strategic, reasons for submitting high ratings.

We do not exhaustively study the determinants of Airbnb's ratings distribution. For instance, social interactions before, during, or after a stay on Airbnb may lead market participants to omit relevant information from their reviews. Furthermore, not all users submit reviews on Airbnb. If those that opt out of reviewing have lower quality experiences, reviews on the platform will tend to be more positive. Our principal stratification results demonstrate this selection. It is also possible that reviewers leave different types of feedback when they know their name and account will be publicly associated with review text. There is room to explore mechanisms that allow reviewers to opt out of associating their review with their profile.

The ratings distribution is also influenced by platform enforcement actions including listing removals and penalties in search rankings. For example, Airbnb's trust and safety team has filtered approximately 970 thousand problematic listings from the platform (Swisher (2019)). We do not know the importance of these actions.

Finally, reviews may describe how an experience compared to the reviewer's own expectations, rather than describing an experience's absolute quality. For example, for cheaper Airbnb listings, guests may not expect hotel quality amenities and service from the host. It should be possible to design review systems that separate expectations-based ratings from more objective evaluations. Indeed, Airbnb has tried to create this separation by asking guests about specific features of a home and grouping listings by those features. "Airbnb Plus" homes not only have high ratings, but are also visited in person by an Airbnb representative to ensure quality, amenities, and the accuracy of the listing description. Similarly, "For Work" homes are those that have WiFi, a work space, and self check-in. The extent to which these complementary reputation mechanisms affect market outcomes remains a question for future work.

References

Airbnb. 2014. "Building Trust with a New Review System – The Airbnb Blog – Belong Anywhere."

Avery, Christopher, Paul Resnick, and Richard Zeckhauser. 1999. "The Market for Evaluations." *American Economic Review*, 89(3): 564–584.

Bolton, Gary, Ben Greiner, and Axel Ockenfels. 2012. "Engineering Trust: Reciprocity in the Production of Reputation Information." *Management Science*, 59(2): 265–285.

Bondi, Tommaso. 2019. "Alone, Together: Product Discovery Through Consumer Ratings."

Brandes, Leif, David Godes, and Dina Mayzlin. 2019. "What Drives Extremity Bias in Online Reviews? Theory and Experimental Evidence."

Cabral, Luís, and Ali Hortaçsu. 2010. “The Dynamics of Seller Reputation: Evidence from Ebay*.” *The Journal of Industrial Economics*, 58(1): 54–78.

Cabral, Luis M. B., and Lingfang (Ivy) Li. 2014. “A Dollar for Your Thoughts: Feedback-Conditional Rebates on Ebay.” Social Science Research Network SSRN Scholarly Paper ID 2133812, Rochester, NY.

Cui, Ruomeng, Jun Li, and Dennis J Zhang. 2019. “Reducing discrimination with reviews in the sharing economy: Evidence from field experiments on Airbnb.” *Management Science*.

Dellarocas, Chrysanthos, and Charles A. Wood. 2007. “The Sound of Silence in Online Feedback: Estimating Trading Risks in the Presence of Reporting Bias.” *Management Science*, 54(3): 460–476.

Ding, Peng, and Jiannan Lu. 2017. “Principal Stratification Analysis Using Principal Scores.” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3): 757–777.

Farronato, A., Andrey Fradkin, B. Larsen, and Erik Brynjolfsson. 2019. “Consumer Protection in an Online World: When Does Occupational Licensing Matter.”

Feller, Avi, Fabrizia Mealli, and Luke Miratrix. 2017. “Principal score methods: Assumptions, extensions, and practical considerations.” *Journal of Educational and Behavioral Statistics*, 42(6): 726–758.

Filippas, Apostolos, John Joseph Horton, and Joseph Golden. 2018. “Reputation Inflation.” *EC ’18*, 483–484. ACM.

Fradkin, Andrey, Elena Grewal, Dave Holtz, and Matthew Pearson. 2015. “Bias and Reciprocity in Online Reviews: Evidence from Field Experiments on Airbnb.” 641–641. ACM.

Frangakis, Constantine E., and Donald B. Rubin. 2002. “Principal Stratification in Causal Inference.” *Biometrics*, 58(1): 21–29.

Hui, Xiang, Maryam Saeedi, and Neel Sundaresan. 2019. “Adverse Selection or Moral Hazard: An Empirical Study.” *Journal of Industrial Economics*.

Jaffe, Sonia, Peter Coles, Steven Levitt, and Igor Popov. 2017. “Quality Externalities on Platforms: The Case of Airbnb.”

Klein, Tobias J., Christian Lambertz, and Konrad Stahl. 2016. "Market Transparency, Adverse Selection, and Moral Hazard." *Journal of Political Economy*.

Lafky, Jonathan. 2014. "Why do people rate? Theory and evidence on online ratings." *Games and Economic Behavior*, 87: 554–570.

Levitt, Steven D., and John A. List. 2007. "What Do Laboratory Experiments Measuring Social Preferences Reveal About the Real World?" *Journal of Economic Perspectives*, 21(2): 153–174.

Livan, Giacomo, Fabio Caccioli, and Tomaso Aste. 2017. "Excess Reciprocity Distorts Reputation in Online Social Networks." *Scientific reports*, 7(1): 3551.

Li, Xinxin, and Lorin M. Hitt. 2008. "Self-Selection and Information Role of Online Product Reviews." *Information Systems Research*, 19(4): 456–474.

Lorenz, Jan, Heiko Rauhut, Frank Schweitzer, and Dirk Helbing. 2011. "How Social Influence Can Undermine the Wisdom of Crowd Effect." *Proceedings of the national academy of sciences*, 108(22): 9020–9025.

Miller, Nolan, Paul Resnick, and Richard Zeckhauser. 2005. "Eliciting Informative Feedback: The Peer-Prediction Method." *Management Science*, 51(9): 1359–1373.

Nosko, Chris, and Steven Tadelis. 2015. "The Limits of Reputation in Platform Markets: An Empirical Analysis and Field Experiment."

Pallais, Amanda. 2014. "Inefficient Hiring in Entry-Level Labor Markets." *American Economic Review*, 104(11): 3565–99.

Proserpio, Davide, Wendy Xu, and Georgios Zervas. 2018. "You Get What You Give: Theory and Evidence of Reciprocity in the Sharing Economy." *Quantitative Marketing and Economics*, 16(4): 371–407.

Swisher, Kara. 2019. "Brian Chesky: How Airbnb is Responding to a Deadly Shooting." *Recode Decode with Kara Swisher*.

Zervas, Georgios, Davide Proserpio, and John Byers. 2015. “A First Look at Online Reputation on Airbnb, Where Every Stay Is Above Average.” Social Science Research Network SSRN Scholarly Paper ID 2554500, Rochester, NY.

A Logging of Reviews

In this section we discuss several details about the logging of review and treatment data in our sample. Overall ratings and review text were required and logged more than 99.9% of the time conditional on a guest review. Whether the other ratings were required depends on the device that was used to submit the review. On iOS, the sub-ratings and recommendations were required. On a desktop browser, the sub-ratings and recommendations were not required and are missing for 5.5% of guest reviews and 4.4% of host reviews. On Android, the sub-ratings were required but the anonymous recommendation was not logged. 79% of guest reviews and 76% of host reviews were submitted via a desktop browser in our sample.

The simultaneous reveal review experiment launched on May 8, 2014 and our sample includes trips with checkout dates between May 7, 2014 and June 12, 2014. However, there were two logging issues during the experiment.

The first logging issue occurred at the outset of the experiment. When launched on May 8, Airbnb's experiment logging framework had bugs. These were fixed by May 11, 2014. Our main analysis sample simply excludes transactions with checkout dates earlier than May 10, 2014. However, if being exposed to the treatment between May 8 and May 11 affected subsequent trips, this could impact our analysis. To verify that this is not the case, we create a new sample that excludes any host with a trip ending on May 7, May 8, or May 9. Note that this sample excludes more active hosts, who are more likely to have a transaction ending on any given day. Figure A4 displays the baseline experimental results for this sample. The treatment effects in the two samples are similar in magnitude and precision.

A second logging issue occurred towards the end of our experiment. Treatment assignment logs are missing for some transactions on June 6 and June 7. We account for this issue with the following procedure. For hosts whose first transaction treatment assignment is missing because it ends on one of these days, we exclude the host from the sample. We keep transactions for hosts whose first transaction is after the June 7 because we can observe treatment assignment.

B Measuring Review Text

The text of a review is the most publicly salient type of the information collected by the review system because the text of a review is permanently associated with an individual reviewer. In this paper, we focus on the sentiment of the text, e.g. whether the text contains only positive information or whether it includes negative phrases and qualifications. We use a regularized logistic regression, a common technique in machine learning, to classify the review text based on the words and phrases that appear in the text.

In order to train a classifier, we need "ground truth" labeled examples of both positive and negative reviews. We select a sample of reviews that are highly likely to be either positive or negative based on the ratings that guests submitted. Reviews by guests that we use as positive examples for guests and hosts are ones that have five star ratings. Reviews by guests that are examples of negative reviews are ones with a one or two star rating. Reviews by hosts that are examples of negative reviews are ones which have either a non-recommendation or a sub-rating lower than four stars. Foreign language reviews were excluded from the sample.

We use reviews submitted between January 2013 and March 2014. Because positive reviews are much more common than negative reviews, the classification problem would be unbalanced if we used the entire sample. Therefore, we randomly select 100,000 examples for both positive and negative reviews. Once we obtain these samples, we remove special characters in the text such as punctuation and we remove common "stop words" such as "a" and "that".[15] Each review is transformed into a vector for which each entry represents the presence of a word or phrase (bigrams and trigrams), where only words that occur at least 300 times are included. We tested various thresholds and regularizations to determine this configuration.

We evaluate model accuracy in several ways. First, we look at the confusion matrix describing model predictions on a 20% hold out sample. For guest reviews of listings, 19% of reviews with low ratings were classified as positive and 9% of reviews with high ratings were classified as

[15] These words are commonly removed in natural language applications because they are thought to contain minimal information.

negative. The relatively high rate of false positives reflects not only predictive error but the fact that some guests misreport their true negative experiences. We also evaluate model accuracy by doing a 10-fold cross-validation. The mean out of sample accuracy for our preferred model is 87%. Figures A1 and A2 display the most common phrases associated with negative reviews by guests and hosts and the relative frequency with which they show up in positive versus negative reviews. Phrases that commonly show up in negative reviews by guests concern cleanliness ('was dirty'), smell ('musty'), unsuitable furniture ('curtains'), noise ('loud'), and sentiment ('acceptable') and phrases that commonly show up in negative reviews by hosts include ('would not recommend'), ('rude'), and ('smoke').

C Experimental Validity

Table AI displays the balance of observable characteristics in the experiments. There are no statistically significant differences in characteristics between the treatment and control guests or listings in the simultaneous reveal experiment.

D Linear Model of Review Timing Effects

In this appendix we discuss an alternative way to study the desire to reveal information using a linear model. We would like to measure the effect of receiving a review on the submission of reviews. In this procedure, the outcome variable is whether a review by a user comes within a day after a review by the counterparty. The sample is the set of of observations for which the focal user (guest or host) does not review first. This includes observations where the focal user does not review at all.

Table AII displays the results from this model. We find that the treatment increases the probability of reviews within a day by 39% for guests and 74% for hosts. These effects persist when adding time of first review fixed effects (columns (2) and (4)). Lastly, in column (5), we interact the treatment with host prior reviews. We do not find substantial or statistically significant heterogeneity in the effect by host experience.

E Principal Stratification Details

In this section, we briefly describe the principal stratification method used to separate the treatment effects we observe into treatment effects on two distinct subpopulations: always takers (i.e., users who would write a review whether in the control or treatment arm of our experiment) and compliers (i.e., users who review their counterparty when enrolled in the treatment, but would not review their counterparty when enrolled in the control). A more detailed description of the principal stratification approach can be found in Ding and Lu (2017).

We first compute the probability that each user in our sample is a complier, always taker, or never taker. We accomplish this by using the marginal method described by Feller, Mealli and Miratrix (2017).[16] Under the principal stratification approach's monotonicity assumption, we can assume that non-reviewers in the treatment group are never takers, and reviewers in the control group are always takers. For all other users in the sample, we can estimate the probability that they are an always taker using a logistic regression model that is trained on data from the control group and predicts the choice to review using a set of user- and trip-level covariates. Similarly, we can estimate the probability that each of these users is a never taker using a logistic regression model that is trained on data from the treatment group and predicts the choice to review using the same set of user- and trip-level covariates. In both cases, we predict the choice to review using the following covariates:

- Whether the guest has any prior trips
- Whether the guest has submitted a review before
- Whether the host has any prior trips
- Whether the host has submitted a review before
- Whether the guest has submitted text before

- The average text sentiment of prior guest reviews
- The average overall star rating of prior guest reviews
- Whether the host has an effective positive percentage (EPP)
- The host's EPP
- Whether the host manages many listings
- Whether the guest has a gender
- Whether the guest has any prior customer service tickets
- Whether the host has any prior customer service tickets
- The property type of the listing
- Whether the guest is from the US
- The log of the listing price
- Whether the booking was made with instant book

Once we have estimated the probability that each user is an always taker and never taker, we can calculate the probability that each user is a complier, since $P(complier)_i = 1 - P(always\, taker)_i - P(never\, taker)_i$. In cases where $P(always\, taker)_i + P(never\, taker)_i > 1$, we set $P(complier) = 0$ and normalize the probabilities that the user is an always taker or never taker so that they sum to 1. After estimating the probability that each user belongs to each stratum, we use these probabilities as weights to construct causal stratum-level treatment effect estimators. Point estimates and confidence intervals are calculated using the bootstrap ($n = 1000$). We use the 'basic' bootstrap confidence interval method from the function 'boot.ci' in R.

[16] We also estimate the probability that each user belongs to each stratum using the EM algorithm described by Ding and Lu (2017). However, in order to make the calculation of bootstrap standard errors computationally tractable, we conduct the majority of our analysis using probabilities obtained through the marginal method. The point estimates we obtain using the EM algorithm are qualitatively similar to those obtained with the marginal method.

We test that the principal stratification model that we have proposed is accurate using the balancing conditions proposed by Ding and Lu (2017). Simply put, the balancing conditions require that within each stratum, the treatment should not appear to have a causal effect on any function of the pretreatment covariates used to estimate a given unit's stratum. We estimate the effect of the treatment on each pretreatment covariate in each stratum. The estimated effects are nearly zero (with a maximum absolute value of 8.07×10^{-7}) across all strata and covariates, indicating that the balancing conditions are satisfied.

F The Long-Run Evolution of Ratings

In this section, we document review rates and the distribution of ratings on Airbnb and how it changes over time. We conduct this exercise to explore the possibility that the longer-run effects of the simultaneous reveal policy may be different from its short-run effects. For example, it may take some time for reviewers to learn about the review system changes or to learn how to best adapt their reviewing behavior given the simultaneous reveal mechanism. If learning or attention were important, we would expect review rates to rise over time and ratings to drop over time as people gradually learn that the review system now prevents retaliation, and adjust their reviewing behavior accordingly. The Airbnb wide launch of the simultaneous reveal policy in July, 2014 may have also increased the attention that users pay to the review system. In this case, we would expect to see changes in reviewing behavior around the time of the launch that are greater than those estimated during the experimental period.

Figure A9 displays the review rates for guests and hosts over time, by treatment group. We see that following the end of the experiment, when all groups were assigned the treatment, the review rates in the control groups quickly jump to match the review rates in the treatment group. This suggests that the longer exposure time for the treatment group did not have first-order consequences for reviewing behavior. It also suggests that the platform-wide launch of the policy did not result in effects larger than those predicted by the experimental treatment effects on review rates.

Figures A10 and A11 display the long-run trends in the distribution of ratings by guests for a set of experienced guests. We focus on experienced guests to minimize the impact of changes in the composition of Airbnb users that may also be occurring during this time period. There are two takeaways from this figure. First, the share of reviews with five stars did drop after the public launch of simultaneous reveal reviews, due to the fact that two-thirds of trips became eligible for the simultaneous reveal system and because of the attention garnered by a blog post and news. Second, the long-run ratings trend did not fall substantially after the initial launch, suggesting that inattention was not a primary driver of the small effects which we found in the experiment.

G Additional Tables

Table AI: Experimental Validity Check

Variable	Difference	Mean Treatment	Mean Control	P-Value	Stars
Total Bookings by Guest	-0.024	2.999	3.024	0.270	
US Guest	-0.002	0.285	0.286	0.558	
Guest Tenure (Days)	-2.065	268.966	271.032	0.271	
Host Listings	0.015	1.858	1.843	0.566	
Listing Reviews	-0.039	10.662	10.700	0.715	
Listing Trips Finished	-0.099	15.091	15.190	0.510	
US Host	0.002	0.266	0.264	0.547	
Multi-Listing	0.002	0.082	0.081	0.262	
Entire Property	-0.001	0.671	0.672	0.682	
Nights	-0.073	5.504	5.577	0.188	
Guests	-0.010	2.360	2.370	0.251	
Price Per Night	-3.138	291.690	294.828	0.273	
Observations	0.001			0.601	

This table displays the averages of variables in the treatment and control groups, as well as the statistical significance of the difference in averages between treatment and control. $p<0.10$, * $p<0.05$, ** $p<0.01$

Table AII: Effects of Treatment on Reviewing Within a Day

	Dependent variable:				
	Reviews Within a Day of First Review				
	Guest		Host		
	(1)	(2)	(3)	(4)	(5)
Treatment	0.024***	0.023***	0.066***	0.064***	0.057***
	(0.002)	(0.002)	(0.003)	(0.003)	(0.006)
Treat * 1 - 3 Reviews					0.013*
					(0.008)
Treat * 4 - 12 Reviews					0.012
					(0.008)
Treat * > 13 Reviews					0.001
					(0.008)
Mean of Y	0.062	0.062	0.089	0.089	0.089
Days Since Checkout FE	No	Yes	No	Yes	Yes
Observations	60,526	60,526	41,563	41,563	41,563
R^2	0.002	0.005	0.014	0.022	0.022
Note:			*p<0.1; **p<0.05; ***p<0.01		

This table displays estimates from a linear probability model where the outcome is whether the guest (columns 1 - 2) or the host (columns 3 - 5) submitted a review within one day after the counterparty. Columns 2, 4, and 5 include fixed effects for the days since checkout of the initial review. ‘Reviews’ in column (5) refer to the number of reviews for the listing at the time of the booking.

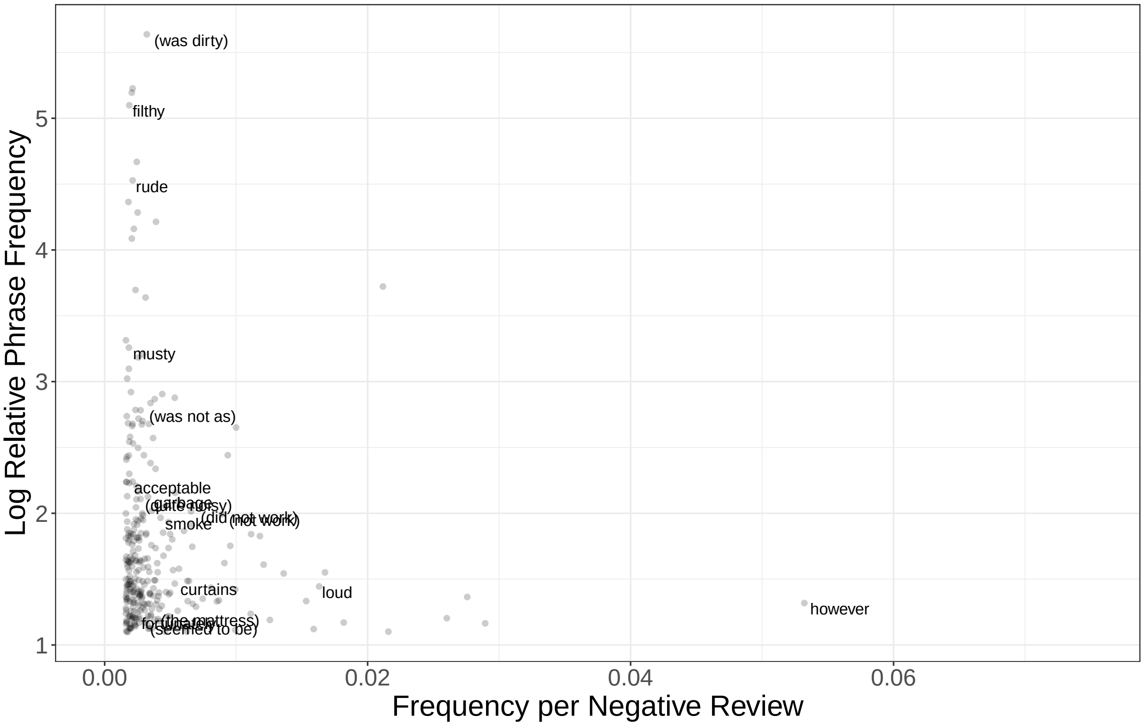
Table AIII: Long-term Guest Outcomes

	Dependent variable:			
	Log(Nights in Exp.)	Log(Trips in Exp.)	Log(Nights by 2015)	Log(Bookings by 2015)
	(1)	(2)	(3)	(4)
Treatment	-0.006	-0.004^{*}	-0.010^{*}	-0.007
	(0.004)	(0.002)	(0.006)	(0.005)
Controls Included	Yes	Yes	Yes	Yes
Observations	115,157	115,157	115,157	115,157
R^2	0.065	0.078	0.186	0.047
Note:				*p<0.1; **p<0.05; ***p<0.01

This regression displays the effects of the treatment on the subsequent Airbnb usage of guests. The outcomes are the log of nights and trips taken during the experimental period as well as the log of nights and bookings which happened before 2015. Controls for guest market of origin, a time trend, the effective positive percentage of the listing, the log of the first price, and the number of reviews of the listing are included. Removing controls does not substantively affect the point estimates.

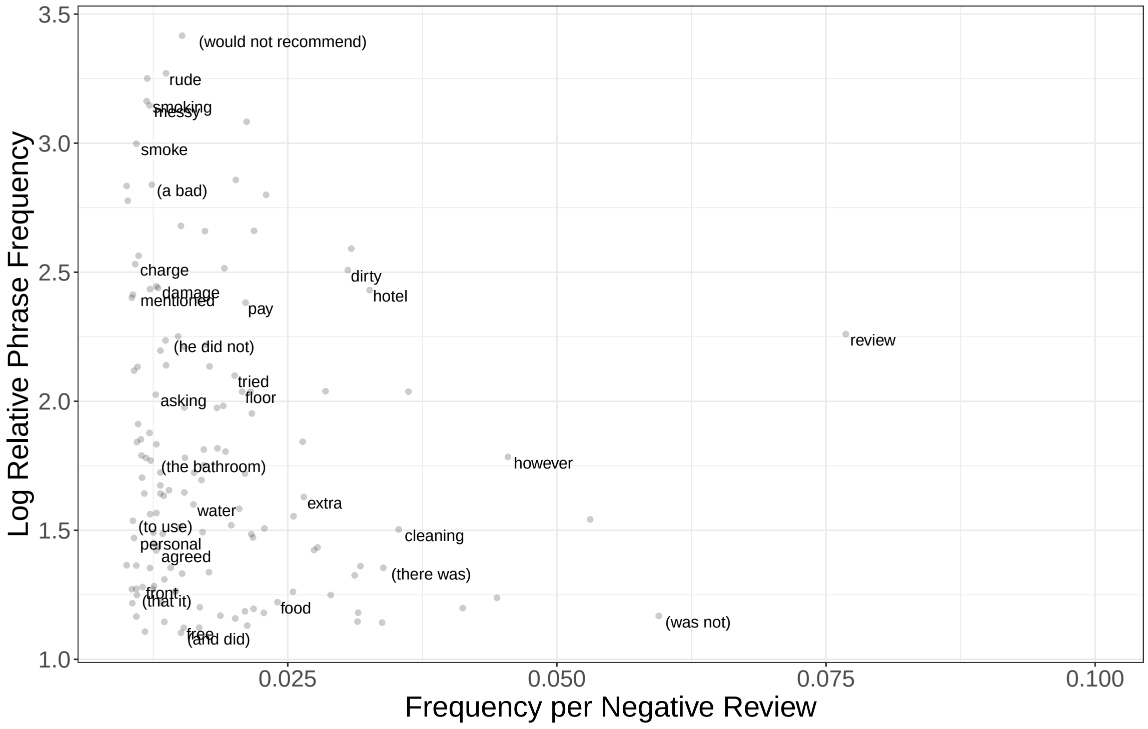
H Additional Figures

Figure A1: Distribution of negative phrases in guest reviews of listings.



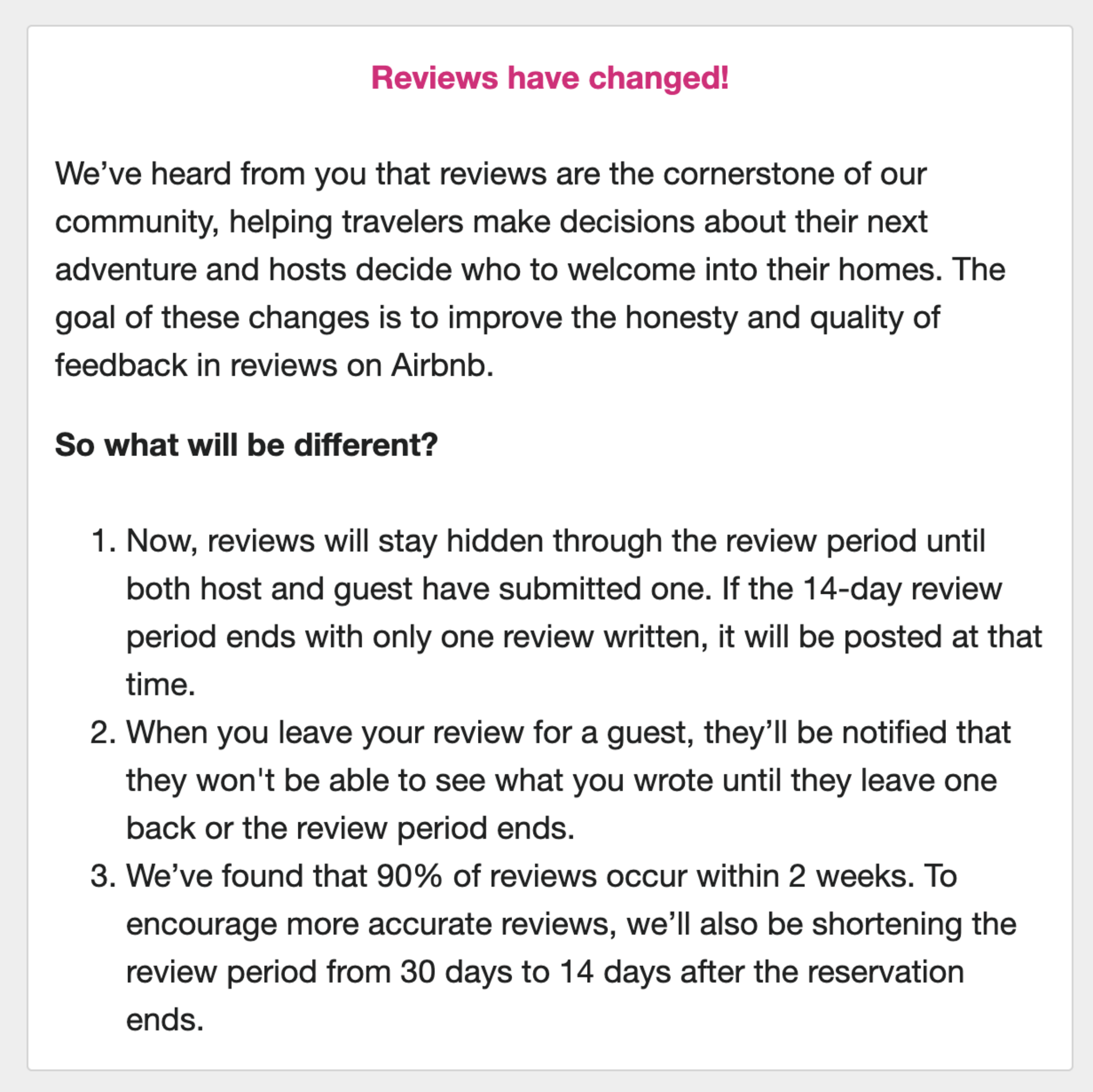
“Relative phrase frequency” refers to the ratio with which the phrase occurs in reviews with a rating of less than five stars.

Figure A2: Distribution of negative phrases in host reviews of guests.



“Relative phrase frequency” refers to the ratio with which the phrase occurs in reviews with a non-recommendation.

Figure A3: Interstitial in Some Host Emails



Reviews have changed!

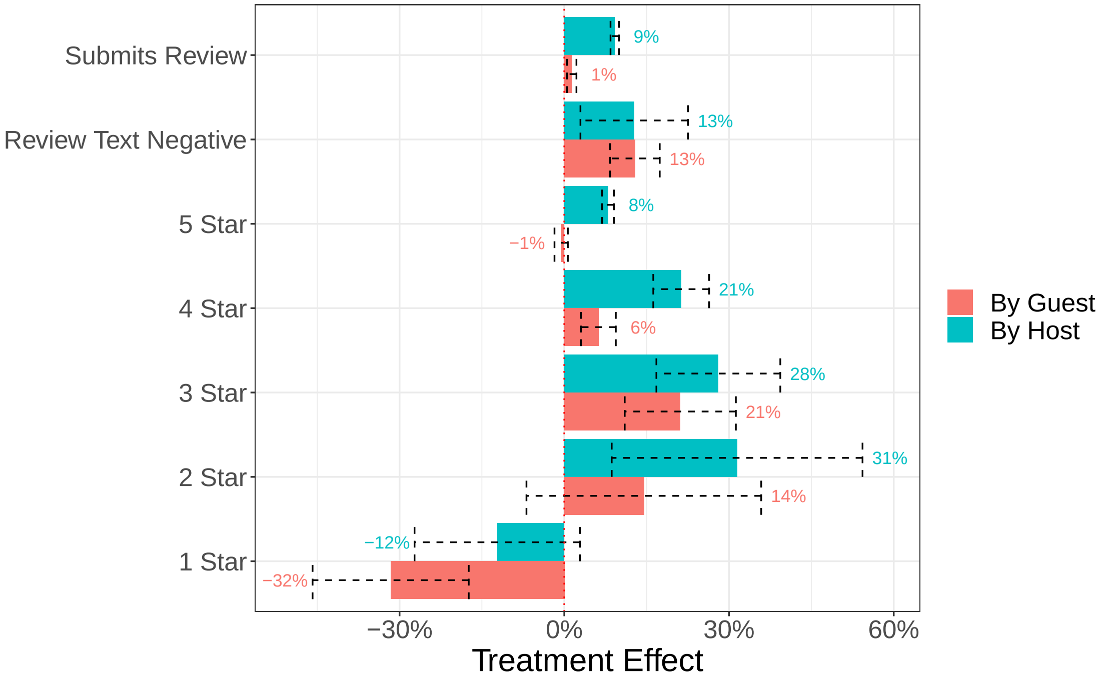
We’ve heard from you that reviews are the cornerstone of our community, helping travelers make decisions about their next adventure and hosts decide who to welcome into their homes. The goal of these changes is to improve the honesty and quality of feedback in reviews on Airbnb.

So what will be different?

1. Now, reviews will stay hidden through the review period until both host and guest have submitted one. If the 14-day review period ends with only one review written, it will be posted at that time.
2. When you leave your review for a guest, they’ll be notified that they won't be able to see what you wrote until they leave one back or the review period ends.
3. We’ve found that 90% of reviews occur within 2 weeks. To encourage more accurate reviews, we’ll also be shortening the review period from 30 days to 14 days after the reservation ends.

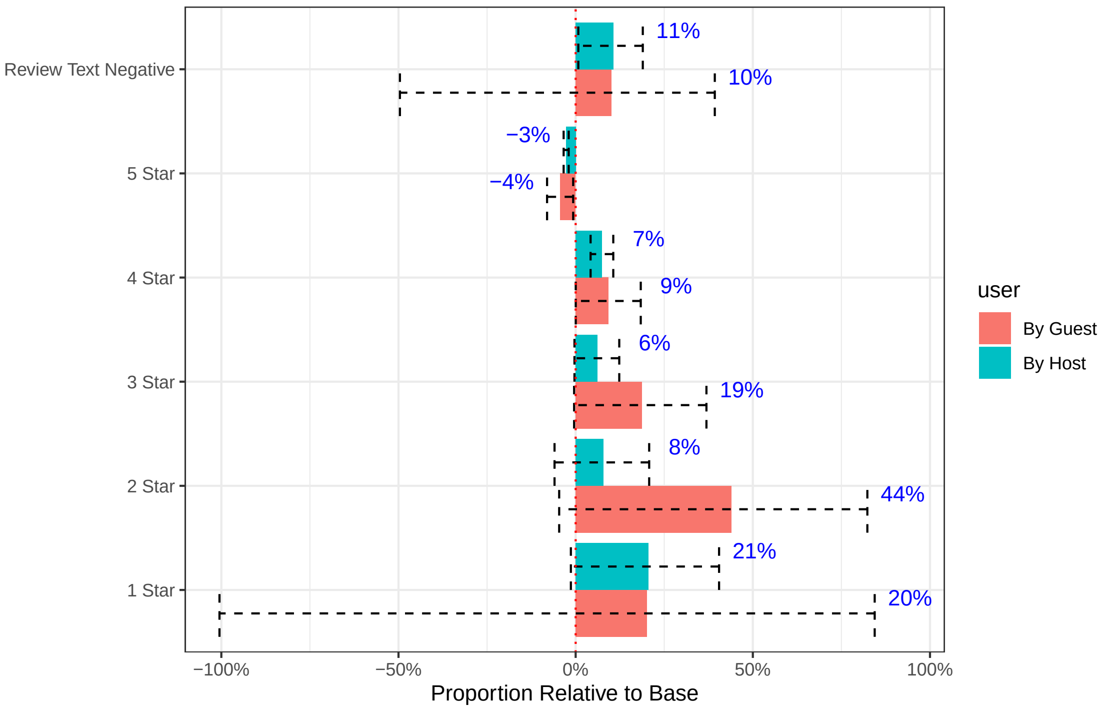
The above figure displays an interstitial inserted into emails received by hosts in the treatment. We are not sure which share of hosts received this interstitial.

Figure A4: Robustness to Alternative Sample: Treatment Effects



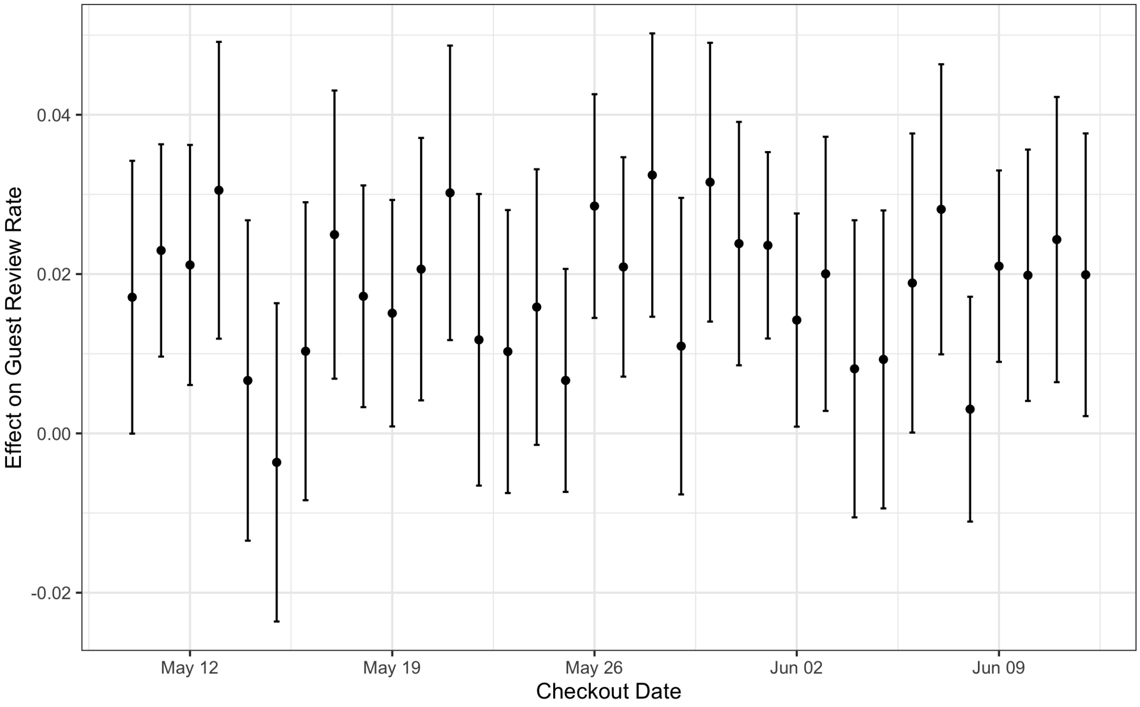
This figure displays the effects on the treatment on reviews by guests and hosts. We measure the percentage effect as the ratio of the absolute treatment effect and the mean in the control. Standard errors used for 95% confidence intervals are computed using the delta method.

Figure A5: Selection Into Reviewing - Complier Causal Effects



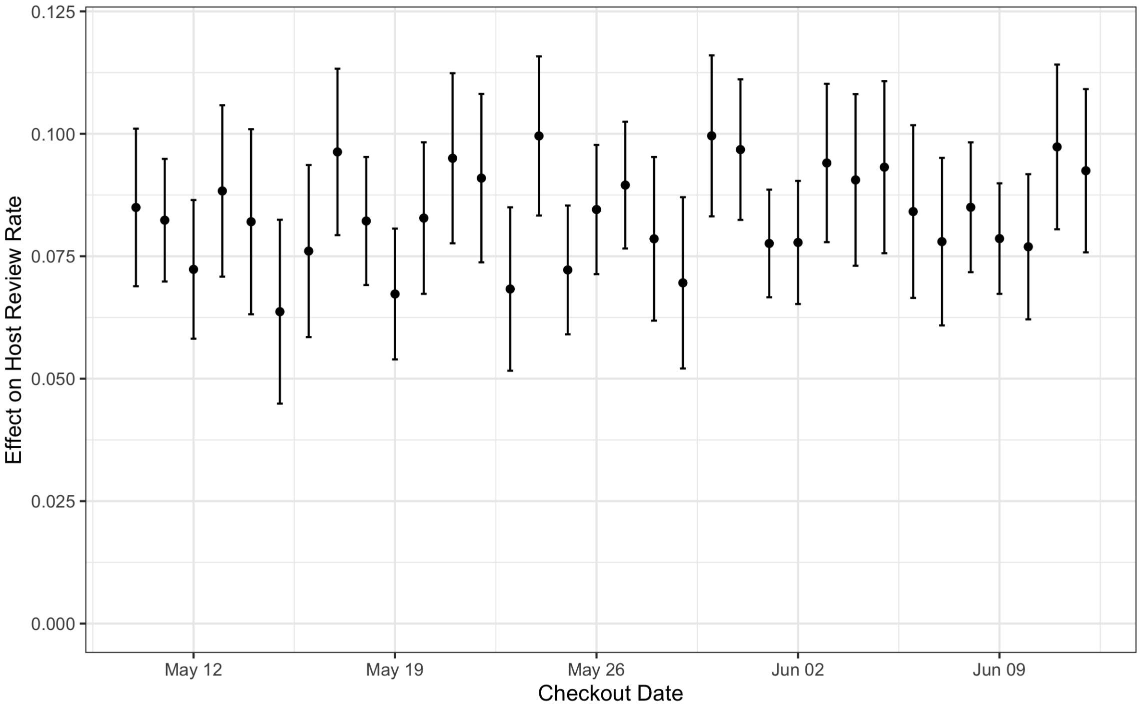
This figure plots the relative likelihood of each review type for compliers (those who review only due to the treatment) relative to the rate for always takers (those who would review regardless of treatment). Confidence intervals are computed using the ’basic’ method.

Figure A6: Effect of Treatment on Guest Review Rates Over Time



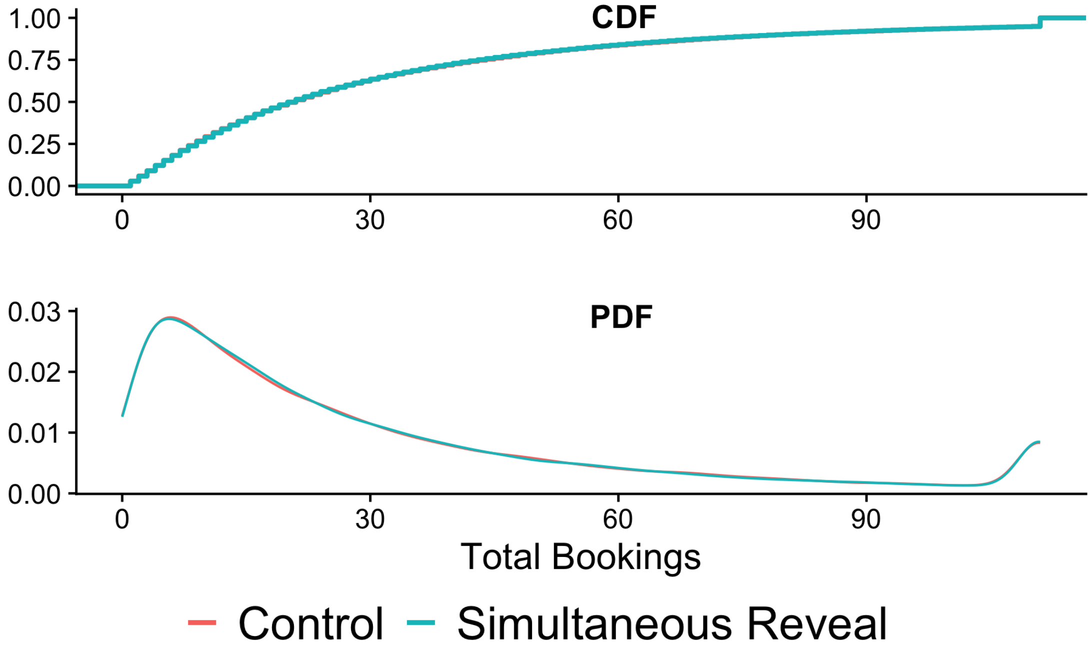
This figure plots the daily treatment effect on guest review rates and the 95% confidence interval. We use all transactions in the sample to increase precision.

Figure A7: Effect of Treatment on Host Review Rates Over Time



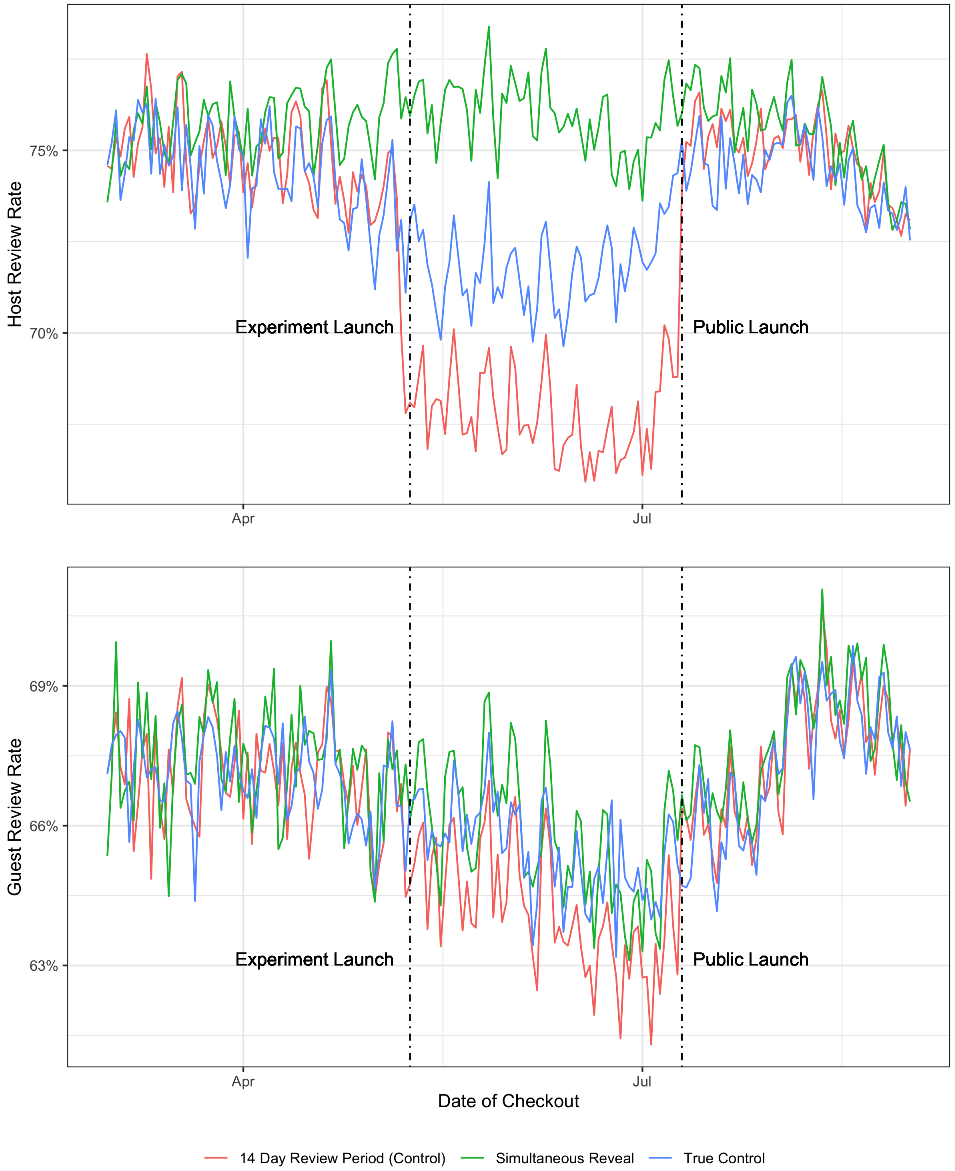
This figure plots the daily treatment effect on guest review rates and the 95% confidence interval. We use all transactions in the sample to increase precision.

Figure A8: Distribution of Bookings by January 1, 2015



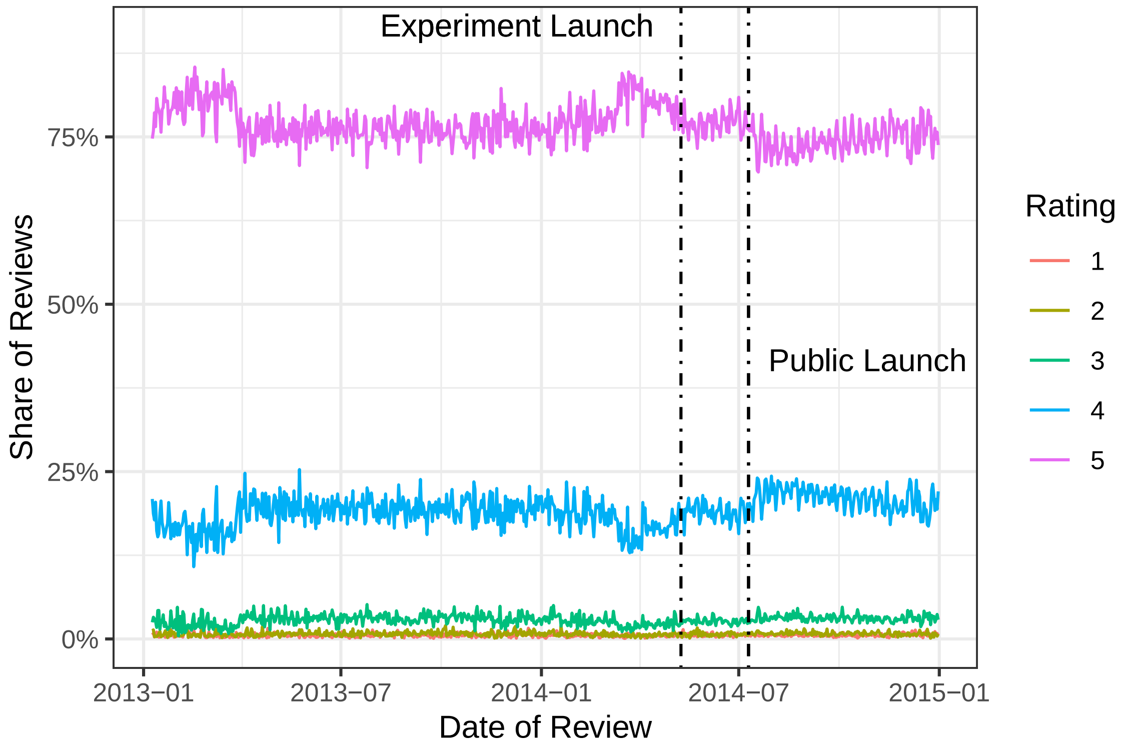
The above figure displays the empirical CDFs and PDFs of total bookings for treated and control listings up to January 1, 2015. We censor the number of bookings at the 95^{th} percentile to make the figure easier to read.

Figure A9: Review Rates Over Time



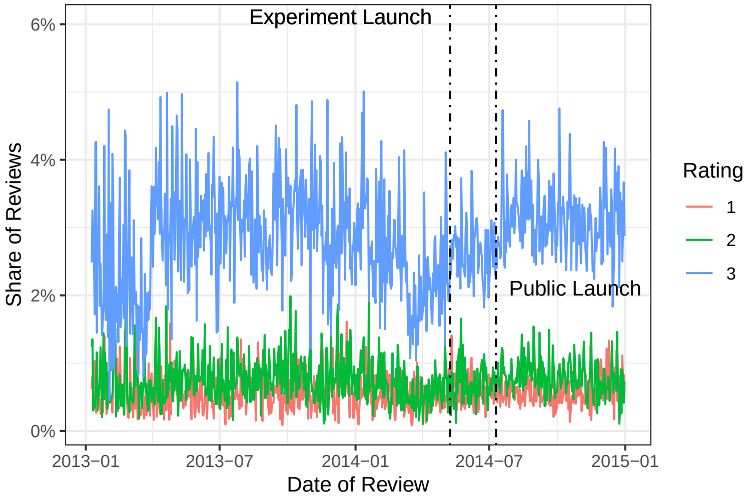
This figure displays the temporal trends of host and guest review rates over time by treatment group. Note that the Simultaneous Reveal Treatment changed the review period to 14 days from 31 days (True Control).

Figure A10: Ratings Over Time



This figure displays the temporal trends of star ratings over time. Because the composition of guests and hosts varies with the growth of the platform, this figure includes only experienced guests reviewing from the domain ("www.airbnb.com") who stayed in a US based listing. The first vertical line demarcates the start of the experiments while the second line demarcates the public launch of the simultaneous reveal system, which placed an additional two-thirds of hosts into the simultaneous reveal treatment.

Figure A11: Ratings Over Time - Low Ratings



This figure displays the temporal trends of star ratings over time. Because the composition of guests and hosts varies with the growth of the platform, this figure includes only experienced guests reviewing from the domain ("www.airbnb.com") who stayed in a US based listing. The first vertical line demarcates the start of the experiments while the second line demarcates the public launch of the simultaneous reveal system, which placed an additional two-thirds of hosts into the simultaneous reveal treatment.