Reciprocity and Unveiling 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]Data 2 The People
[3]MIT Sloan School of Management

December 26, 2020

Abstract

Reputation systems are used by nearly every digital marketplace, but designs vary and the effects of these designs are not well understood. We use a large-scale experiment on Airbnb to study the causal effects of one particular design choice — the timing with which feedback by one user about another is revealed on the platform. Feedback was hidden until both parties submitted a review in the treatment group and was revealed immediately after submission in the control group. The treatment stimulated more reviewing in total. This is due to users' curiosity about what their counterparty wrote and/or the desire to have feedback visible to other users. We also show that the treatment reduced retaliation and reciprocation in feedback and led to lower ratings as a result. The effects of the policy on feedback did not translate into reduced adverse selection on the platform.

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 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 an upwardly biased reputation system can cause buyers to unexpectedly transact with low-quality sellers, which in turn makes them less likely to transact on that platform again 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 change to Airbnb's reputation system. The system is two-sided, meaning that the guest and the host each have the opportunity to review one another. In the control group of our experiment, reviews are revealed both to the counterparty and to the public as soon as they are submitted. 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 simultaneous reveal treatment changes the timing with which a review is revealed to the counterparty and on the platform. In the treatment group, reviews 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 for users in the treatment group to retaliate against a negative review. 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 on the subsequent demand for treated sellers.

The treatment reduced the time to review by 17% for guests and 9.9% for hosts. It also increased review rates by 1.7% for guests and 9.8% for hosts. We hypothesize that these treatment effects are largely driven by a mechanism that has not previously been documented, which we refer

to as the *review unveiling mechanism* — users want to see and reveal what the other user wrote. This may be caused by simple curiosity, or by a strategic desire to have information revealed more quickly to potential trading partners. In support of the review unveiling mechanism, we show that while the treatment decreased the time from checkout to first review 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. In concordance with predictions from prior literature (Bolton, Greiner and Ockenfels (2012)), 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. The treatment also decreased the correlation between guest and host ratings by 48%.

Next, we consider whether the lower ratings in the treatment represent more accurate ratings due to the reduction in reciprocity, or whether they are solely due to changes in who reviews. Since the treatment increased review rates, 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. This, in addition to our results regarding the correlation of guest and host ratings, provides evidence that the effects on review ratings reflect more accurate reviews in the treatment group.

Lastly, we consider the effects of the treatment on subsequent host outcomes on the platform. If the reputation system became more informative due to simultaneous reveal, then treated sellers, 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 listing demand. 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 test for heterogeneous effects across seller types and find no evidence that ex-ante worse hosts are hurt by the treatment. Our findings contrast with Bolton, Greiner and Ockenfels (2012) and Hui, Saeedi and Sundaresan (2019), who find that similar changes to reputation systems decreased demand for low quality sellers. We attribute this contrast to a number of

3

factors, which we discuss 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 and section 7 we discuss treatment effects and evidence regarding causal mechanisms. Section 8 contains the results of robustness checks and section 9 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 in two-sided markets. 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 peo-

ple 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 in two-sided markets. 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]

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 when mutually negative feedback occurs, the second review occurs quickly after the first. This is stated as evidence for 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 review rates, ratings, 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

---

[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) find evidence of reciprocity in content platforms.

proposed policy (Levitt and List (2007)). On the other hand, as we discuss in section 8, 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. In particular, the incentive to have reviews revealed quickly is an important driver of our results, and 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 both 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 for 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. (2020). 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. (2020) show how transactions with low quality sellers reduce guests' subsequent usage of the Airbnb platform.

3   Theoretical framework

The game of reciprocal reviewing with variable review timing has not, to our knowledge, been formalized in the preceding literature. Prior work has instead made informal arguments about the effects of reciprocal feedback. Namely, that positive feedback induces positive feedback in response and that negative feedback triggers retaliation. In this section, we add an additional mechanism to the theory of reciprocal reviewing, which we call the *review unveiling mechanism*. The explanation for this mechanism is that the desire to reveal information about a review provides an incentive for agents to review more often and more quickly in the simultaneous reveal (SR) reputation system. In the subsequent empirical sections, we argue that prior theories of reciprocal reviewing are not sufficient by themselves, and that this additional mechanism helps to explain our

results.

We begin by summarizing the arguments in the prior literature. Prior work is implicitly based on a utility function consisting of the following terms.

1. An intrinsic cost (or benefit) from leaving a review. This will vary across individuals so that some people dislike reviewing while others like it.

2. A dis-utility from misrepresenting the quality of a transaction in a review. This implies that absent other forces individuals review honestly.

3. For a second review, there is a positive utility from submitting a review with a reciprocal rating. For example, the second reviewer may feel obliged to leave a positive review once they read the positive review of the first reviewer.

4. The reviewer cares about her own reputation. Since the first review affects the rate and type of second reviews, the reviewer rationally takes this into account. This leads to more positive ratings and fewer negative ratings in a first review relative to the situation when only 1) and 2) operate.

We can now compare the utility of reviewing in the SR versus the review in turn (RIT) reputation system. We discuss the implications separately for the second reviewer and then for the first reviewer. Throughout the discussion below we assume that the opportunity to review first or second is exogenous.[2]

The second reviewer knows that her review will not change anything about the first review. As a result, term 4 drops out under both systems. In the SR system, the second reviewer does not know the content of the first review. As a result, she is not affected by the content of the first review, and term 3 drops out. In contrast, in the RIT system, term 3 remains. Since term 3 increases the utility of reviewing, the utility of reviewing is lower in the SR system and the review rate is predicted to fall. This is the argument made by Bolton, Greiner and Ockenfels (2012) and confirmed in their laboratory experiment.

---

[2]There is also an incentive in RIT to wait until the other party reviews in order to threaten retaliation. Our data does not suggest that this is an important motivation on Airbnb. Hosts typically review first and more positively than guests. This is true even though hosts have much more to lose from a negative review than guests (Cui, Li and Zhang (2019)). We interpret this as evidence that most hosts value the benefit of inducing a positive review more than the benefit of waiting to threaten a negative review.

The first reviewer in the SR system knows that the content of the first review will not affect the content of the second review. As a result, the first reviewer has less reason to review and to do so positively in SR than in RIT. Consequently, if the only mechanism driving reviewing behavior was reciprocity, then first reviewer's review rate would be predicted to fall as well.

We now add one more term into the utility function, which corresponds to the review unveiling mechanism. Our theory states that the presence of a 'hidden' first review that can be revealed increases the utility of reviewing. Suppose you are a host who has received an email notifying you of a new guest review. Naturally, you would like to know what the guest said, but you can't until you review the guest. Furthermore, if you expect the review is positive, you might also want it displayed publicly as quickly as possible, so that you can receive more bookings. This combination of curiosity and strategic behavior motivates you to leave a review right away, rather than wait.

While it is standard to take strategic considerations into account when studying markets, the role of curiosity has less frequently been considered. The information conveyed in a review is similar to gossip and other social information, which is the topic of much of human conversation and has been shown to trigger curiosity (Dunbar, Marriott and Duncan (1997)). More generally, curiosity has been shown to strongly affect behavior (see: Loewenstein (1994) and Silvia (2012)). Given the similarities between online reviews and other sources of social information, we predict that curiosity should be present in the setting of reviews.

The key prediction of the review unveiling mechanism is that there will be more reviews and faster reviews in SR. In particular, the second review should arrive faster in SR than in RIT. This prediction is opposite to the prediction yielded by the standard reciprocity motive.

The review unveiling mechanism may also have an effect on the speed of first review. Namely, users may want to write a quick first review in order to trigger a quick second review. We posit that the effect on first review timing will be smaller than the effect on second review timing, since the existence of a treatment effect for first review timing depends on first reviewers understanding that a quick first review will trigger a quick second review. Given that we had to write this research paper, it is reasonable to assume that many first reviewers have not considered this implication of

9

the SR system.

To conclude the theoretical framework section, we consider the effects of switching to SR on ratings. The removal of the ability to condition the second rating on the first in SR should result in less inflated ratings and less correlation between reviews (Bolton, Greiner and Ockenfels (2012)). The review unveiling mechanism does not change that prediction. Even in the SR system, there may still be some reciprocity. For example, the host may give a negative rating in anticipation of a negative rating from the guest. Nonetheless, reviews in SR should be less influenced by reciprocity and should be more correlated with the underlying quality of the transaction (Klein, Lambertz and Stahl (2016); Hui et al. (2014)). If reviews are more informative, then worse listings are less likely to be booked, or will be forced to lower prices, i.e., SR should reduce adverse selection. We measure the magnitude of the effects on adverse selection in section 9.

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.[3] 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
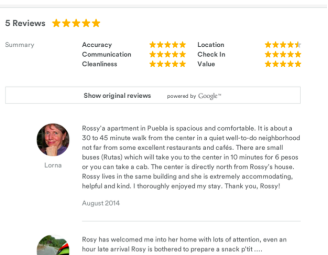
cannot change their reviews after they have been submitted.

At the time of the simultaneous reveal experiment, 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,[4] and private comments to the host.
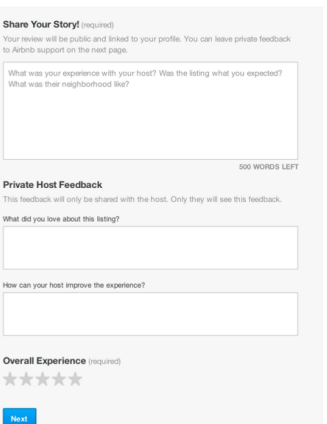
The next page asked guests to rate the listing in six specific categories: accuracy of the listing compared to the guest's expectations, the communication of the host, the cleanliness of the listing, the location of the listing, the value of the listing, and the quality of the amenities provided by the 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 a 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.[5]

Figure 1: Review flow on the website
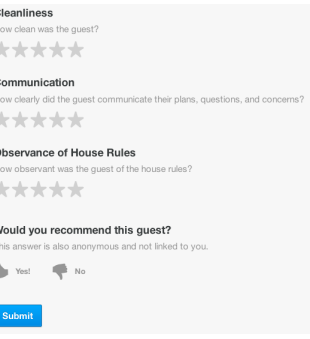
(a) Reviews on Listing Page **(b)** Review of Listing (Page 1) **(c)** Review of Guest (Page 2)

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

---

[3]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.

[4]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.

[5]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 of the simultaneous reveal experiment and reviewing patterns in the control group. 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.[6] 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 from 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

---

[6]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 ████,

Please leave a review for your recent host, ███████████. 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.

Leave a review

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 ███,

Thought you'd like to know that your recent host ████████, 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.

Leave a review

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

Figure 3: Control Email - Guest

(a) First Email

Hi ████,

We hope you enjoyed your stay with ████████ residence in Taksim district. Please help ████████ 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 ████████.

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 ████████,

You have received a review from ████████! 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 Appendix F, we discuss the potential impact of these changes on our results. Finally, both guests and hosts received a prominent notification before starting a review (Figure 5).
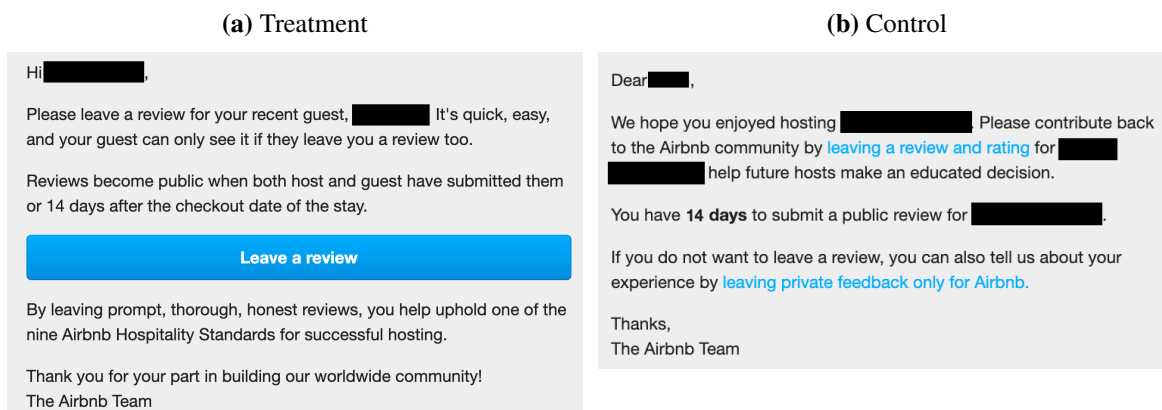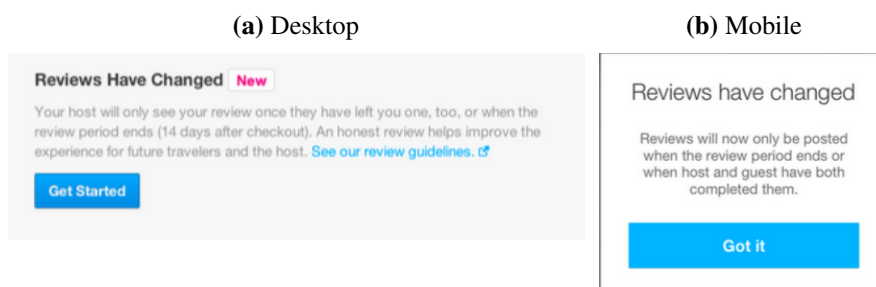
Figure 4: Host First Emails

(a) Treatment

Hi ▮▮▮▮▮▮,

Please leave a review for your recent guest, ▮▮▮▮▮▮ 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.

Leave a review

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 ▮▮▮▮,

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

You have **14 days** to submit a public review for ▮▮▮▮▮▮.

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.

Get Started

(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.

Got it

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

Below, we describe reviewing behavior in the 14 day control. Throughout sections 5, 6, and 7, we focus on the first transaction observed for each host either in the treatment or in the control.[7] We later turn to the effects of the experiment on subsequent reviews and stays.

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.[8] On average, users review frequently and positively, with hosts reviewing more positively and faster than guests. 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. The average time between a first and second review in the control group is 3 days. This is an important statistic for testing the review unveiling mechanism, and we will return to it in section 6.

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 86% of host reviews containing five star ratings for all categories.

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

---

[7]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, which may affect subsequent transactions and reviews. Since guests and hosts in this sample did not know about the change to the review system before the trip, they cannot adjust their match to the new policy. In contrast, for subsequent transactions, the treatment may affect selection into transactions. Furthermore, this sample restriction allows us to avoid issues due to spillovers between multiple listings managed by the same host.

[8]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

(a) Ratings Distribution

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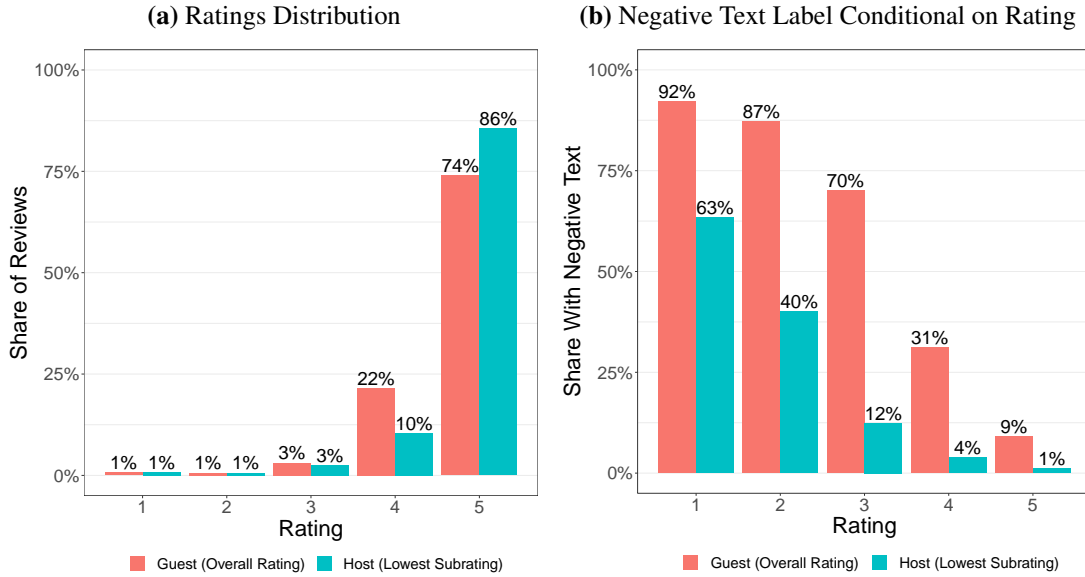
The left figure displays the distribution of submitted overall ratings by guests and lowest category ratings by hosts in the control group of the simultaneous reveal experiment. The right figure displays the prevalence of negative text review as predicted by a regularized logistic regression conditional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

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. 90% of one to 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 31% of the time for four star reviews and 9.2% of the time for 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 measurement error in our text classification procedure.

6 The review unveiling mechanism

In this section, we provide experimental evidence in support of the review unveiling mechanism. As discussed in section 3, if reviewing is driven by the desire to reveal information, then the SR treatment should increase review rates and the speed of reviews, particularly following a first review. In contrast, if the main effect of SR is to reduce reciprocity, then we would expect review rates to fall.

Table 1 shows the control and treatment means as well the treatment effect for review timing related variables. Review rates increase by 1.7% for the guest and by 10% for the host. Importantly for the unveiling mechanism, the number of days between reviews falls by 35%. This is much larger than the fall in the overall time to review (17% for guests and 9.7% for hosts).

Table 1: Summary Statistics

	Control Mean	Treatment Mean	Effect
Submits Review (Guest)	0.68	0.69	0.01 ***
Submits Review (Host)	0.72	0.79	0.07 ***
Days to Review (Guest)	4.70	3.89	-0.81 ***
Days to Review (Host)	3.80	3.42	-0.37 ***
Days Between Reviews	3.05	1.98	-1.07 ***
Days to First Review	3.33	3.01	-0.32 ***

This table displays mean outcomes in the control and treatment, as well as treatment effects. 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;

The large drop in the time between reviews suggests that the first review triggers a much faster second review in the SR treatment, as predicted by the review unveiling mechanism. We now test for this formally by modeling the time to review as a duration. In particular, we expect that both in the treatment and in the control, a first review increases the hazard of the second review. This occurs because the first review automatically triggers an email sent to the counterparty, which reminds the other user to review. Furthermore, we predict that the first review has a larger effect on second review hazard in the SR treatment, because the second reviewer wants to reveal the review.

Our empirical specification is displayed below and represents the canonical Cox proportional hazards model.

$$\lambda_i\left(t \mid \mathbf{x}_i\right) = \lambda_0(t)\exp\left\{\mathbf{x}_i'\boldsymbol{\beta}\right\} \tag{1}$$

In the above equation, $\lambda_i(t)$ is the hazard rate of reviewing for individual i at time t. Our covariates, x_i, include an indicator for the SR treatment, an indicator for whether the time is after the counterparty (guest or host) has reviewed, and an interaction between treatment and being the the first review. We are interested in both the baseline effect of the treatment on reviewing and the interaction term.

Table 2 displays estimates of Cox proportional hazard models of review hazards for guests (columns (1) and (2)) and hosts (columns (3) and (4)). We find that the treatment increases the overall hazard of reviews for guests by 8% (column 1) and hosts by 23% (column 3) — which is consistent with the fact that the review unveiling mechanism causes faster reviews.

Table 2: Speed of Review Effects

	Dependent variable:			
	Relative Hazard of Review			
	Guest		Host	
	(1)	(2)	(3)	(4)
Treatment	1.080	1.014	1.230	1.108
	t = 11.003***	t = 1.450	t = 31.003***	t = 12.560***
After Host		2.470		
		t = 86.862***		
Treat * After Host		1.123		
		t = 8.276***		
After Guest				1.989
				t = 63.448***
Treat * After Guest				1.554
				t = 30.948***
Number of Events	82055	82055	90034	90034
R^2	0.001	0.084	0.006	0.089
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.

A key prediction of our theory is that the SR treatment should cause an especially large effect

on the speed of the second review relative to the first. This is because the second review instantly reveals the first review. To test for this, we interact the treatment with whether the counterparty has already reviewed (columns (2) and (4)).[9]

We find that a first review increases the hazard of a second review in both the treatment and the control.[10] The interaction effect between the treatment and a first review is 12% for guest second reviews of hosts and 55% for host second reviews of guests. Both of these interaction effects are statistically significant.

For both guests and hosts, we find that the effect of the treatment on reviewing is mostly explained by this interaction. To see this, compare the coefficient on the treatment in columns (1) and (2). It falls from 1.08 to 1.01, meaning that the hazard of guest reviews does not increase in the treatment until a host leaves a review. Similarly, when comparing columns (3) and (4), the baseline effect of the treatment on the hazard rate falls from 1.23 to 1.11, meaning that the treatment mostly increases host reviewing through the increased speed of a second review.

The above evidence is consistent with a large effect of the review unveiling mechanism. Not only do reviews rates increase, but the hazard model shows that faster second reviews after an initial first review explain most of the total effect. We conclude that the review unveiling mechanism is salient when the second review immediately reveals the contents of the first review.

7   Reciprocity and its effects on ratings

The above section demonstrated that, contrary to the predictions of a model with only reciprocity, the simultaneous reveal (SR) treatment caused review rates to increase. This means that the review unveiling mechanism was more influential than the reciprocity mechanism in determining review rates. We now show that the treatment effects on ratings are consistent with a decrease in

---

[9]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 AIII).

[10]The fact that even in the control group a first review speeds up the second may be explained by one of three factors. First, the first review may serve as a reminder. Second, the first review may induce a reciprocal obligation to review. Lastly, the speed of guest and host reviews may be correlated with each other due to unobserved heterogeneity.

reciprocity in the SR treatment.

Recall that because SR eliminates the ability to reciprocate the rating of a first review, ratings should decrease and should be less correlated between the guest and host. Figure 7 displays the treatment effects split by the star rating submitted (we do not condition on whether a review is submitted). Consistent with the first prediction, we see a pronounced increase in reviews with 2- to 4-star ratings. On the other hand, reviews with 5-star ratings don't increase for guests and increase to a much lesser extent for hosts.

Figure 7 also documents a fall in 1-star ratings in the treatment. We posit that this effect is also due to reduced reciprocity. Namely, when negative ratings occur in the control group, they often trigger 1-star retaliatory reviews. Since the SR treatment prevents guests from seeing the first review content, we observe fewer 1-star ratings. One supporting piece of evidence for this explanation is that the fall in 1-star ratings by guests is particularly large (66%) for cases in which the first review contains negative text. We see a similar pattern for host reviews after a negative guest review. That said, 1-star reviews are rare; there are just 290 1-star reviews in the control group and 201 1-star reviews in the treatment group.

Figure 7: Effects of Experiment on Reviews
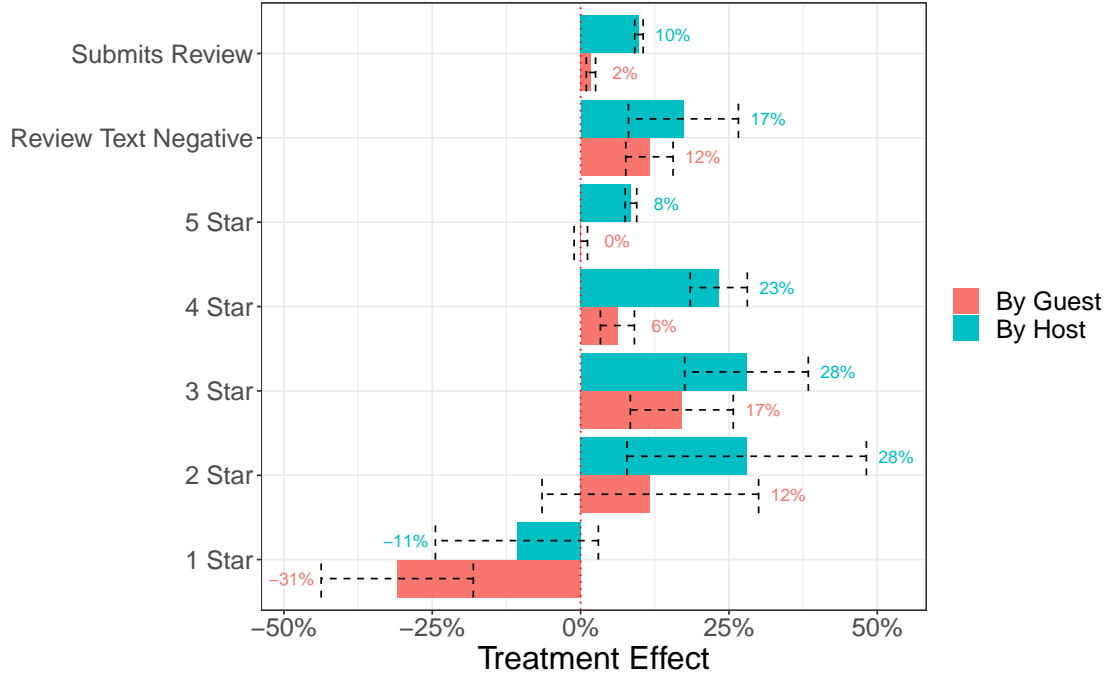
This figure displays the percentage change (relative to the mean in the control group) in reviews of a given type due to the treatment. The standard errors used for the 95% confidence intervals are calculated using the delta method. Transactions with no label, such as when there is no review, are treated as zeros for the purpose of this calculation.

Next, we test the prediction that review content between guests and hosts should be less correlated as a result of the SR treatment. We measure the correlation in reviews across two different measures: the labeled review text, and the lowest rating (including sub-ratings). 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%).

In summary, the changes in the observed ratings and the fall in the correlation between guest and host ratings are consistent with the simultaneous reveal treatment reducing reciprocity. In the next section, we consider alternative explanations for our empirical findings, including whether the effects on ratings are caused by changes in who reviews, rather than reductions in reciprocal behavior among those who do review.

8   Alternative explanations of experimental effects

In this section, we consider alternative explanations for the treatment effects of the simultaneous reveal policy. We focus on two main threats to our interpretation. First, it could be the case that increases in review rates are caused by unintended changes in Airbnb's review solicitation emails. Second, it could be that changes in the submitted ratings are caused by changes in who reviews, rather than in changes to how people review. We discuss both of these threats below and relegate additional robustness concerns to Appendix F.

8.1   Do unintended changes in the email explain increases in review rates?

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 email text (i.e. 'Thank you for your part in building our worldwide community!'). It could be the case that these confounding changes - and not reciprocity or the reveal unveiling mechanism - explain the treatment effects we observe. Below, we argue that these changes are unlikely to explain the treatment effects.

Consider the fact that both the first and second treatment emails sent to users have a large blue button. Furthermore, the first email has text asking the reviewer to be 'prompt and honest'. If these email features cause increases to review rates, then we would expect them to matter more for the first email sent by Airbnb, since these are presumably seen first by the user. Instead, we find that the effects of the treatment are largest for the second review (Table 2). In fact, our hazard models show that for guests, the treatment effect on review rates only shows up after the host has submitted a review. This is inconsistent with the simple theory that a larger button causes substantially more reviews.

A related point is that there are other ways in which users may learn about the treatment policy and these are not affected by the confounding email text. Users are alerted to the new policy not only in the review email, but also during the review flow. Anyone logged in on Airbnb.com or on the app is also shown an alert asking them to submit a review. In summary, while we acknowledge

that confounding changes to the email text may have effects, we believe they are unlikely to explain the increased review rates we observe.

Lastly, there was some variation in the email text sent to the treatment group over the course of the experiment. We believe these variations in the email text do not undermine our tests of the review unveiling mechanism, and provide further evidence for this argument in Appendix F.

8.2 Does selection into who reviews explain the fall in average ratings?

We now consider an alternative explanation for the observed changes to ratings in the treatment. Dellarocas and Wood (2007), Fradkin et al. (2015), and Brandes, Godes and Mayzlin (2019) all argue that who selects into reviewing affects rating distributions. Since the simultaneous reveal treatment increased review rates, it could be the case that changes in ratings are explained by a change in the composition of reviewers, rather than a fall in reciprocity as argued in section 7. We use the methodology of principal stratification (Frangakis and Rubin (2002) and Ding and Lu (2017)) to show that the observed changes in ratings are not solely caused by changes in the selection of reviewers.

Principal stratification is a procedure for identifying the effects of a treatment for latent sub-groups in the experimental sample. The effects on these subgroups provide insight into the causal mechanisms underlying the overall treatment effects. In our setting, we posit that there are three latent types of individuals:

- *Always reviewers*. These individuals review regardless of whether they are in SR or the control.

- *Compliers*. These individuals are induced to review by the SR treatment and would not review if they were in the control condition.

- *Never reviewers*. These individuals never review.

Any effect of the treatment on always reviewers is, by definition, free from selection. The method of principal stratification by principal scores (Ding and Lu (2017)) allows us to estimate this treatment effect. The key assumption required for implementing principal stratification is

23

called weak general principal ignorability. It states that the expected outcome, conditional on submitting a review, is independent of latent strata (complier and always reviewer) when controlling for covariates.[11] 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.

The procedure is conducted in several steps. We first use a logistic regression trained on data from the control group to predict the choice of whether to review as a function of user- and trip-level covariates. Similarly, we use a logistic regression in the treatment group to predict the decision to not review using the same covariates. Once we have these probabilities, we can calculate the probability that (conditional on covariates) a user is a never reviewer, always reviewer, or complier. Finally, we can use a weighting procedure to calculate the stratum-specific causal effects. We discuss the details of this procedure in Appendix E.

To evaluate the fit of our predictive models, we consider our ability to predict reviewing behavior out of sample. We use a 10-folds cross-validation procedure. This procedure produces out of sample predictions that we use to calculate the area under the curve (AUC) of the ROC and generate a calibration plot for these predictions. For host reviews of guests, we achieve an AUC of 0.74 while for guest reviews of hosts, we achieve a lower AUC of 0.64. Both of these AUC measurements are better than the null of no-predictive power. The predictions are also well calibrated (Figure A1).

Using the principal stratification approach, we find that the treatment *does* change reviewing behavior for the always reviewers — 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 reviewers are caused to submit more 2- to 4-star ratings relative to 1- or 5-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 also how

---

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

Airbnb users reviewed their counterparty conditional on leaving feedback.[12]

Figure 8: Always Reviewer Causal Effects
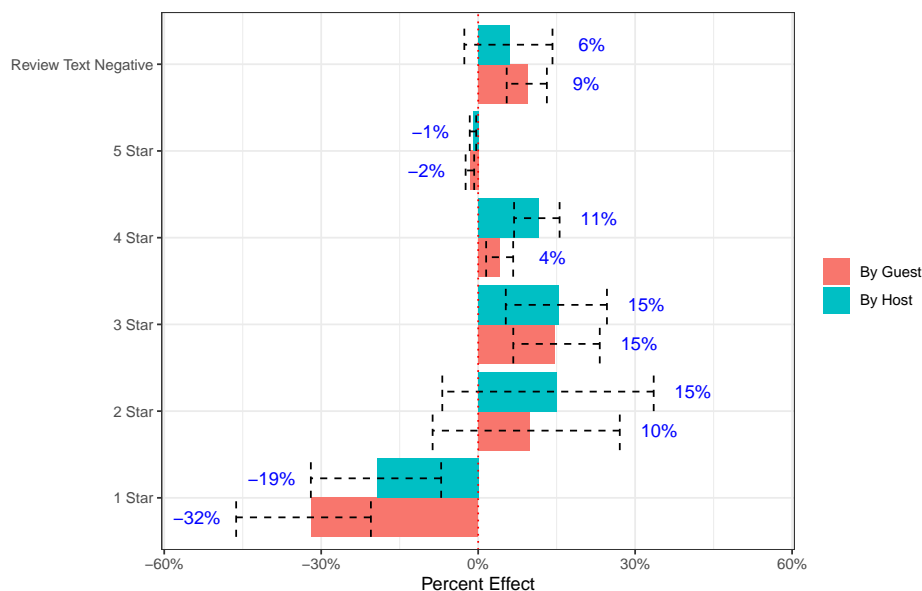
This figure displays the percentage change (relative to the mean in the control) in reviews of a given type due to the treatment. The standard errors used for the 95% confidence intervals are calculated using the percentile bootstrap method. Transactions with no label, such as when there is no review, are treated as zeros for the purpose of this calculation.

One concern about the principal stratification procedure is that it assumes monotonicity, which may be violated if the absence of reciprocity in the treatment causes some individuals to not review. We follow Ding and Lu (2017) in testing for the robustness of our results to violations of the monotonicity assumption. We call individuals who would review in the control but not in the treatment defiers, following the standard terminology in settings of experimental non-compliance. We assume that that number of defiers is 33% of the number of compliers, and recompute the always reviewer causal effects (Figure A3). We find very similar results, showing that modest violations of monotonicity do not overturn our findings.

In summary, we've shown that the effects of the simultaneous reveal on ratings are not caused by changes in who selects into reviewing. Instead, they are caused by changes in the ability of reviewers to condition ratings on the content of the first review.

---

[12]The composition of ratings submitted by compliers is displayed in Figure A2. The ratings left by compliers are typically lower than those of always reviewers.

9 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 studying 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.[13]

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.[14] 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 review unveiling mechanism. 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 the 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 close 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

---

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

[14]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
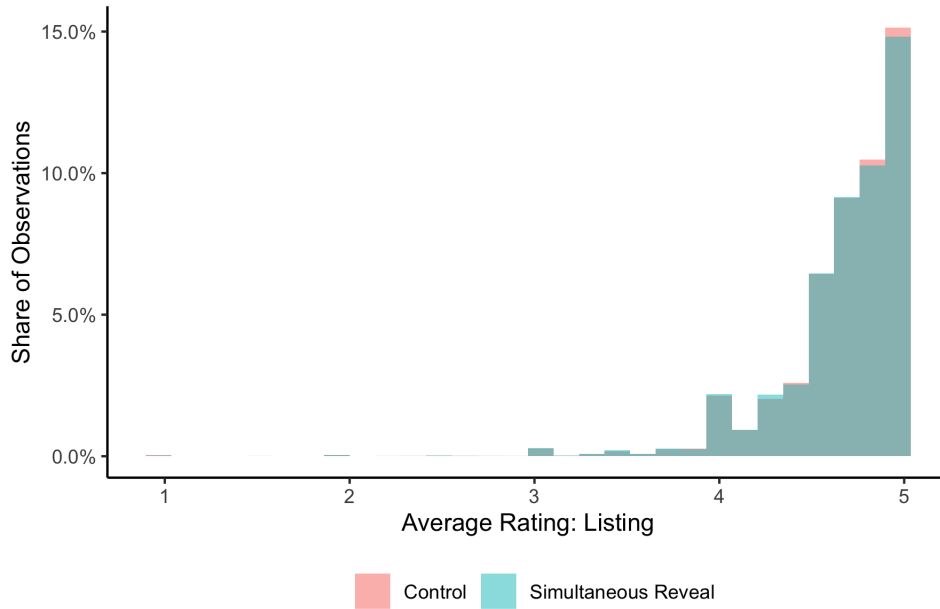
This figure displays the distribution of the average rating across reviews within a listing following the first transaction in the experiment.

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.[15] Such a decrease in demand for ex-ante worse listings 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.

---

[15]Another possibility is that the treatment reduces moral hazard, which we were unable to test for since we cannot

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.[16] 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.

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

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.

[16]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.

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 A4 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.

10   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. To our knowledge, all major platforms with two-sided review systems have implemented systems where users are unable to see their counterparty's review before writing their own. However, reviews are unveiled to the reviewer only on some platforms (Upwork and Freelancer) but not on others (Lyft and Uber). In this paper, we study the effects of a simultaneous reveal policy intended to reduce reciprocity and to improve market outcomes. Our results suggest that whether the review is unveiled plays a critical role in the effects of the simultaneous reveal design.

We find that the simultaneous reveal policy increased review rates and decreased the average valence of reviews. It also reduced retaliatory 1-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 what we refer to as the review unveiling mechanism. 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 review unveiling mechanism 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 period of time prior to simultaneous reveal reviews being launched across the entire site.

We draw several other 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 show that this can be explained by the review unveiling mechanism, 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

31

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 that who selects into reviewing can effect the observed rating distribution. 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 970,000 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.

Dunbar, Robin IM, Anna Marriott, and Neil DC Duncan. 1997. "Human conversational behavior." *Human nature*, 8(3): 231–246.

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*.

Hui, Xiang, Shen Shen, Maryam Saeedi, and Neel Sundaresan. 2014. "From Lemon Markets to Managed Markets: The Evolution of eBay's Reputation System."

Jaffe, Sonia, Peter Coles, Steven Levitt, and Igor Popov. 2020. "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.

Loewenstein, George. 1994. "The psychology of curiosity: A review and reinterpretation." *Psychological bulletin*, 116(1): 75.

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."

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.

Silvia, Paul J. 2012. "Curiosity and motivation." *The Oxford handbook of human motivation*, 157–166.

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

Yarkoni, Tal. 2019. "The Generalizability Crisis."

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 A5 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".[17] 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

---

[17]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 A6 and A7 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 and Additional Results

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.

Table AII displays the control and treatment means, as well as the treatment effects, for the following review related variables: whether the overall rating was 5 stars, whether all category ratings were 5 stars, and whether the review text was classified as negative.

D Linear Model of Review Timing Effects

In this section we discuss an alternative way to study the review unveiling mechanism 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 AIII 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 reviewers (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 reviewer, or never reviewer. We accomplish this by using the marginal method described by Feller, Mealli and Miratrix (2017).[18] Under the principal stratification approach's monotonicity assumption, we can assume that non-reviewers in the treatment group are never reviewers, and reviewers in the control group are always reviewers. For all other users in the sample, we can estimate the probability that they are an always reviewer 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 reviewer 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 reviewer and never reviewer, we can calculate the probability that each user is a complier, since $P(complier)_i = 1 - P(always\,r)_i - P(never\,r)_i$. In cases where $P(always\,r)_i + P(never\,r)_i > 1$, we set $P(complier) = 0$ and normalize the probabilities that the user is an always reviewer or never reviewer 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.

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 Additional concerns with the experimental results

F.1 Changes in email text over the course of the experiment

One concern with our experiment is that the exact email copy sent to users varied over the course of the experiment. We do not have internal data about which user got which email and even about the universe of emails sent.

To investigate whether these changes in email text were important, we solicited Airbnb review emails from this time period via social media. For guests in the control group, we found three versions of the email, which varied in the color scheme and logo.[19] 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.

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 A8) and was deployed for some members of the

---

[18]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.

[19]Airbnb introduced a new logo and color scheme on July 16, 2014, which is after our experiment concluded.

treatment group.

We know that Airbnb was not randomizing the specific email copy concurrently with our experiment. Therefore, any changes to the email copy must have occurred over time, with a change on a particular date. To test whether these changes in the email copy were material, we estimate the effects of the treatment across the days of our experiment. The results are displayed for guests in Figure A9 and hosts in Figure A10. In both cases, there are no trends in the treatment effect over time and the effect is similar in magnitude across the days. This confirms that any changes to the email copy during our experiment did not have large effects on the treatment effect.

As a final comment, to the extent that the email copy changed during our experiment, our treatment effects reflects a mix of the email copy that was sent to different users. This, if anything, improves the external validity of our estimates since there are many ways a platform could inform users about a new reviewing policy and other platforms may do so in a way which is more similar to one or the other email sent by Airbnb.[20]

F.2 Learning about the treatment

One concern with our interpretation of the experimental treatment effects is that users may not immediately learn about the change to the reputation system. 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. This would attenuate the effects that we detect in our sample, but would not reverse our findings regarding the review unveiling mechanism and reduced reciprocity. We provide evidence that learning effects were not of first order importance.

Figure A11 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 and the review rates of the treatment group do not jump. Therefore, the longer exposure time for the treatment group did not have first-order consequences for reviewing rates. This shows that

---

[20]See Yarkoni (2019) for a discussion about the importance of using multiple stimuli for generalization.

learning by users over the course of the experiment did not substantially affect review rates. 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.

In A12 and A13 we plot how ratings 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. Namely, the share of reviews with five star ratings falls and the share of reviews with 3 and 4 star ratings increases. Our results about both the review rates and ratings shows that the potential of learning over time about the policy does not overturn our main results.

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: Other Experimental Treatment Effects

	Guest			*Host*		
	Control Mean	Treatment Mean	Effect	Control Mean	Treatment Mean	Effect
Overall Rating = 5	0.74	0.73	-0.01 ***			
All Ratings = 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	1.8e-03 *

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

Table AIII: 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 AIV: 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: Calibration plot of review prediction in control group

(a) Host Reviews of Guest

(b) Guest Reviews of Host
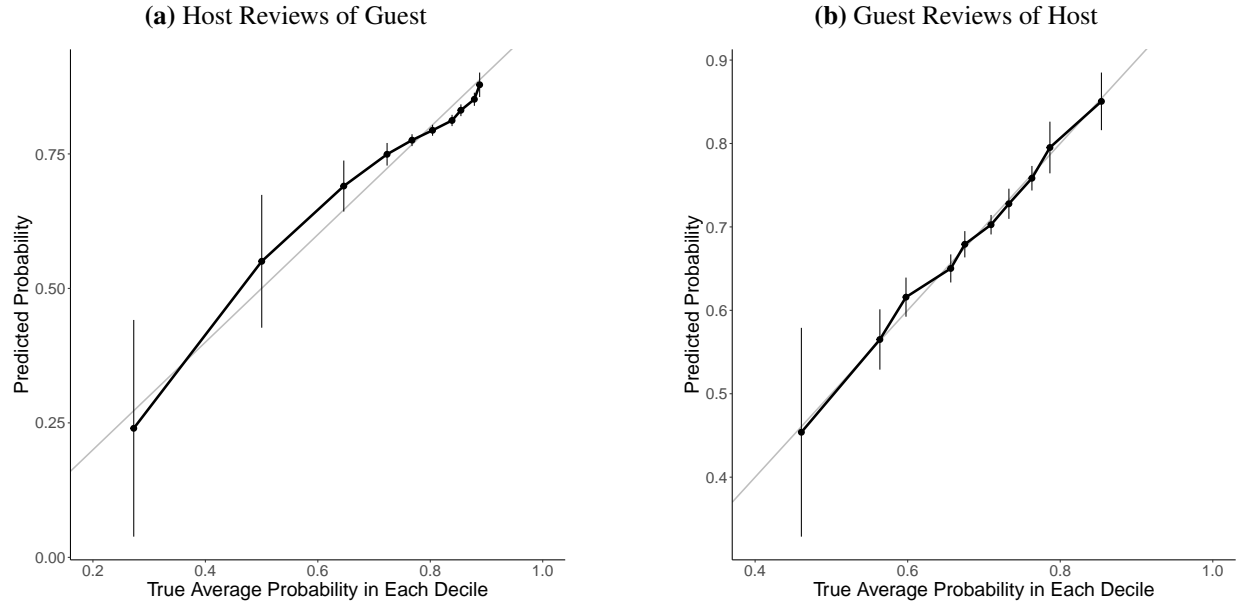
This figure plots the average review probability in the control group against the model predicted probabilities (y-axis). The line range represents the 95% range of predicted probabilities. The data is grouped into deciles of predicted probabilities so that each bin has approximately the same number of observations. The model used for prediction is described in Appendix E and 10-folds cross-validation is used to make the prediction out-of-sample.

Figure A2: Selection Into Reviewing - Complier Causal Effects
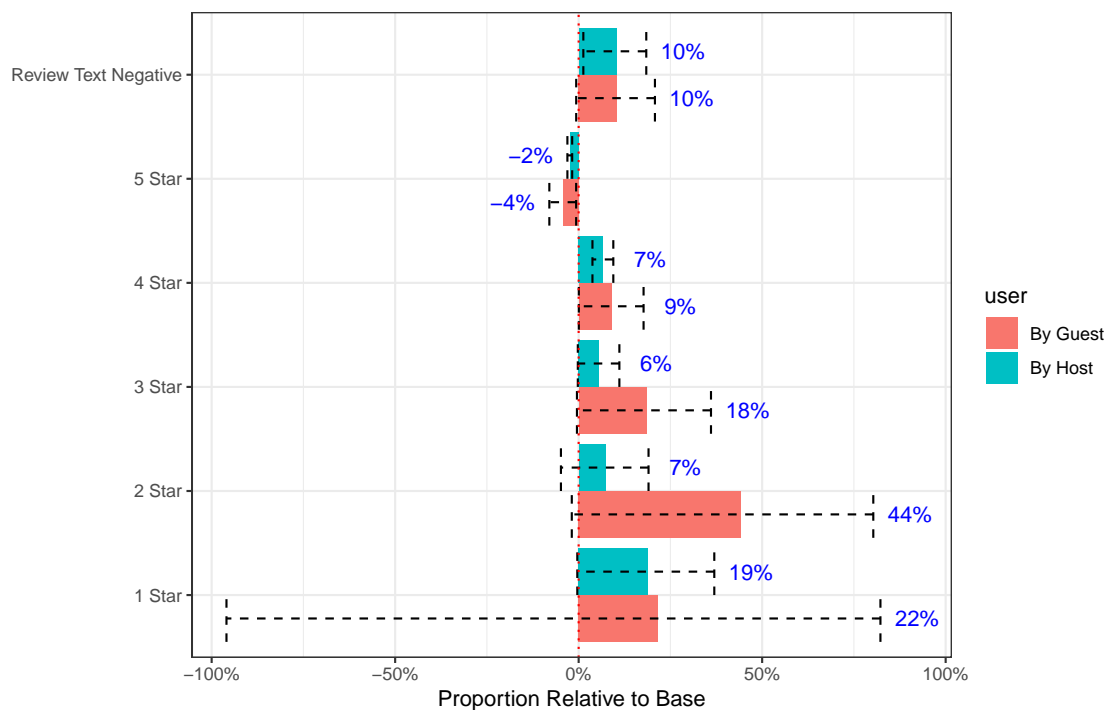
This figure plots the relative likelihood of each review type for compliers (those who review only due to the treatment) relative to the rate for always reviewers (those who would review regardless of treatment). Confidence intervals are computed using the 'basic' method.

**Figure A3:** Always Reviewer Causal Effects - Monotonicity Robustness
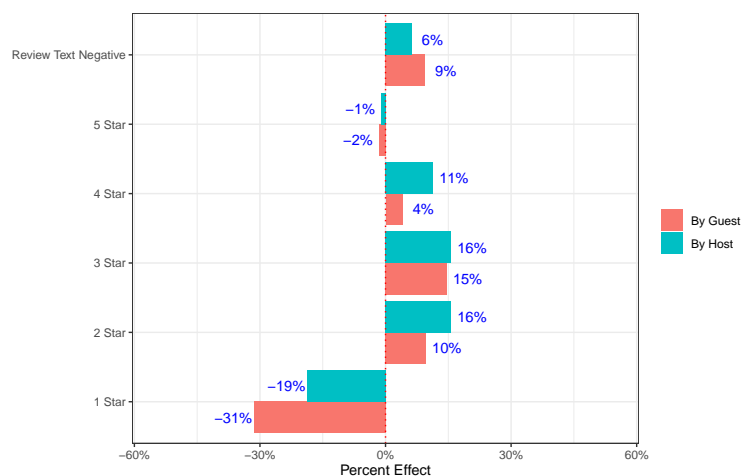
This figure displays the estimated effects of the treatment on always reviewers with the assumption that there are defiers (those who review in the control but not the treatment). We assume that the number of defiers is 33% the number of compliers (those who review in the treatment but not in the control).

**Figure A4:** Distribution of Bookings by January 1, 2015
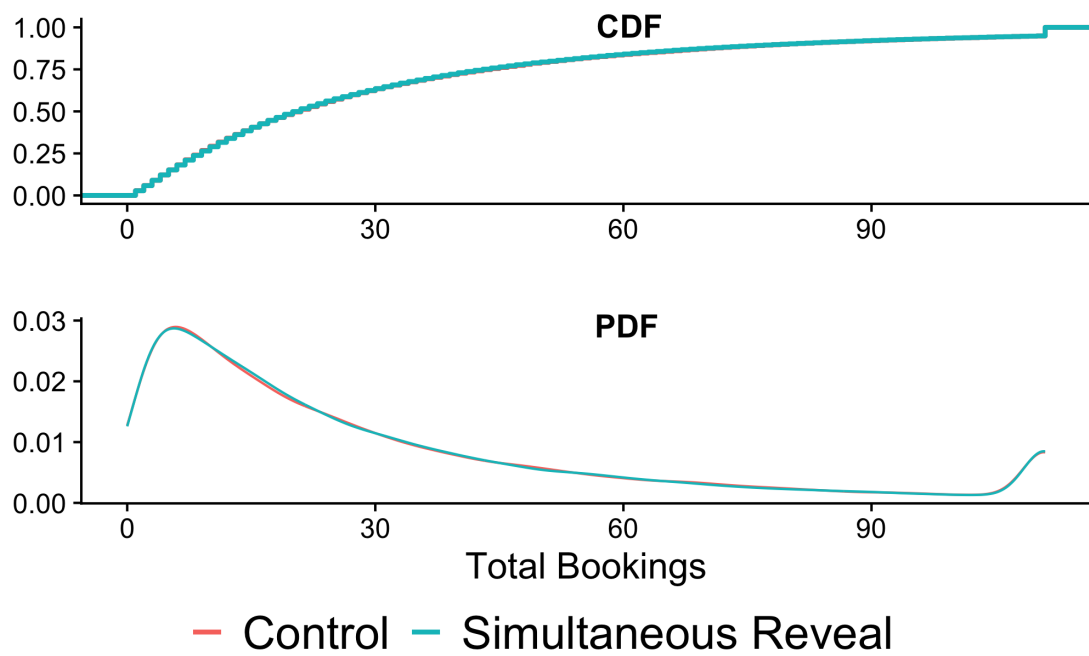
The above figure displays the empirical CDFs and PDFs of total bookings for treated and control listings up to January 1, 2015. We censor the number of bookings at the 95th percentile to make the figure easier to read.

Figure A5: Robustness to Alternative Sample: Treatment Effects
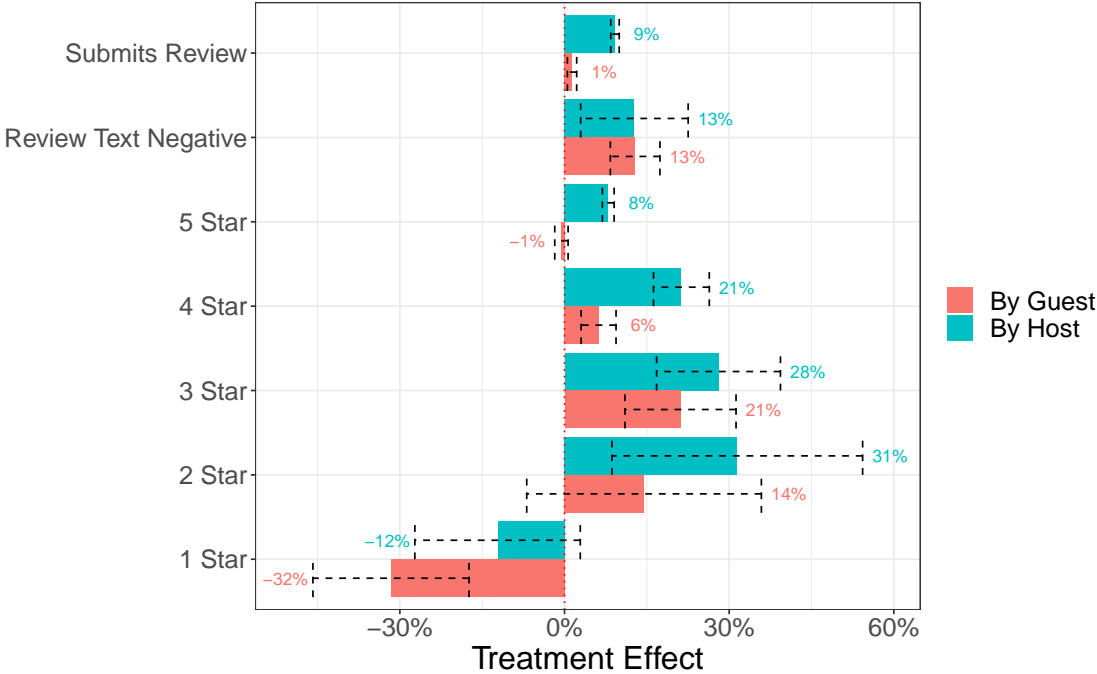

This figure displays the effects on the treatment on reviews by guests and hosts. We measure the percentage effect as the ratio of the absolute treatment effect and the mean in the control. Standard errors used for 95% confidence intervals are computed using the delta method.

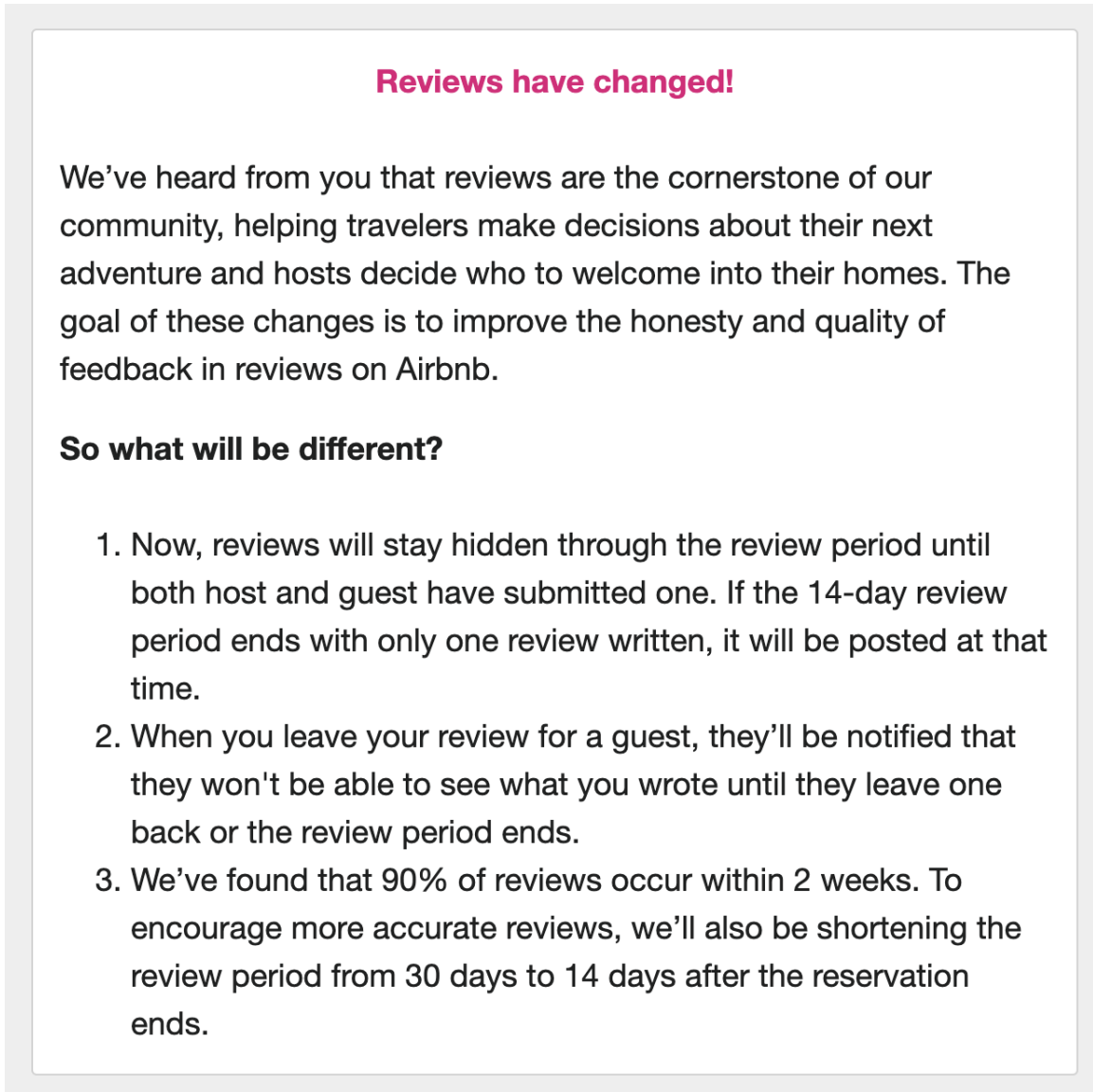**Figure A6:** Distribution of negative phrases in guest reviews of listings.
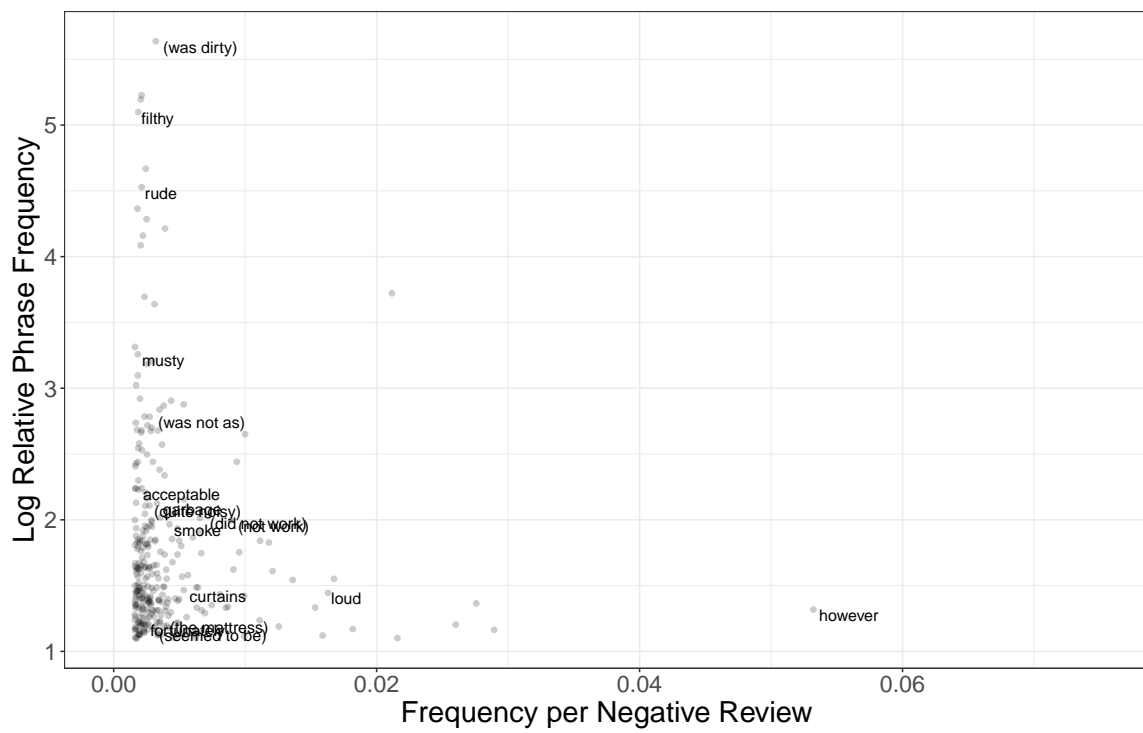
"Relative phrase frequency" refers to the ratio with which the phrase occurs in reviews with a rating of less than five stars.

**Figure A7:** Distribution of negative phrases in host reviews of guests.
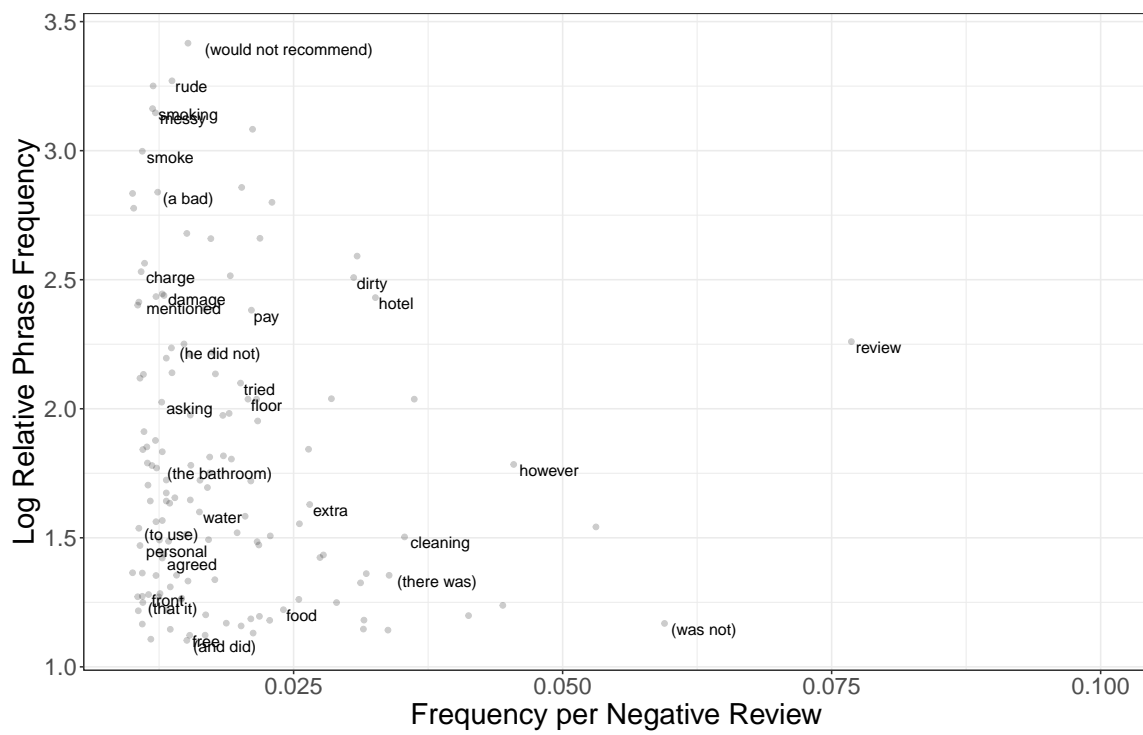
"Relative phrase frequency" refers to the ratio with which the phrase occurs in reviews with a non-recommendation.

Figure A8: 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 A9:** Effect of Treatment on Guest Review Rates Over Time
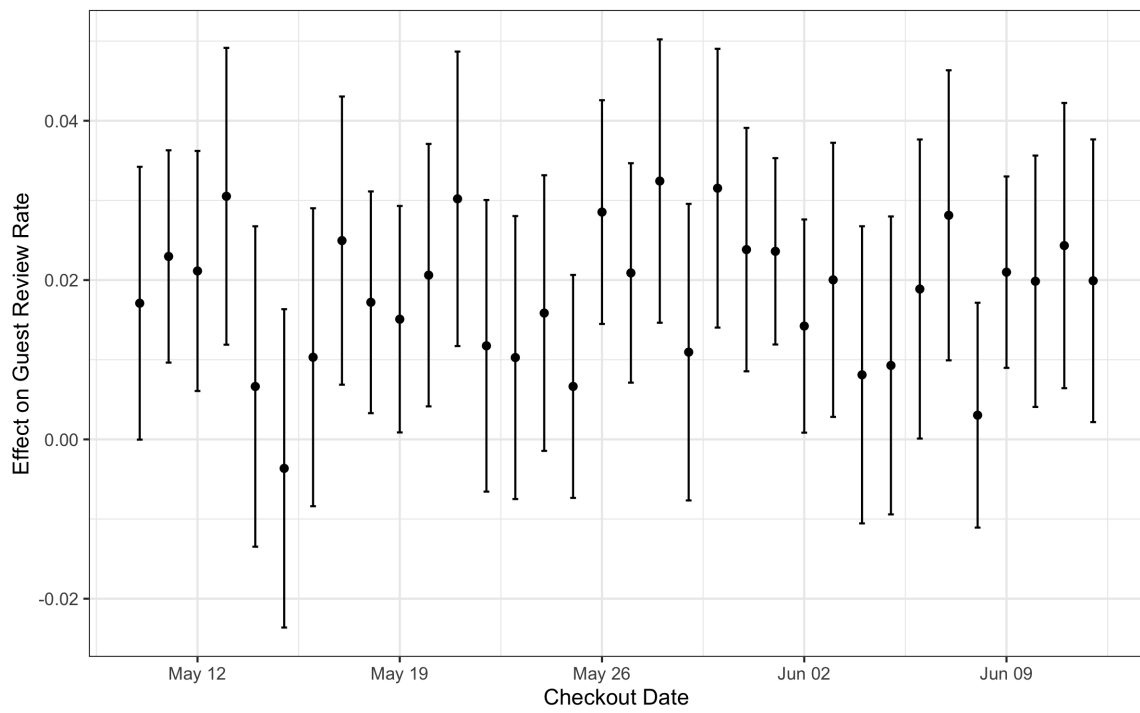
This figure plots the daily treatment effect on guest review rates and the 95% confidence interval. We use all transactions in the sample to increase precision.

Figure A10: Effect of Treatment on Host Review Rates Over Time
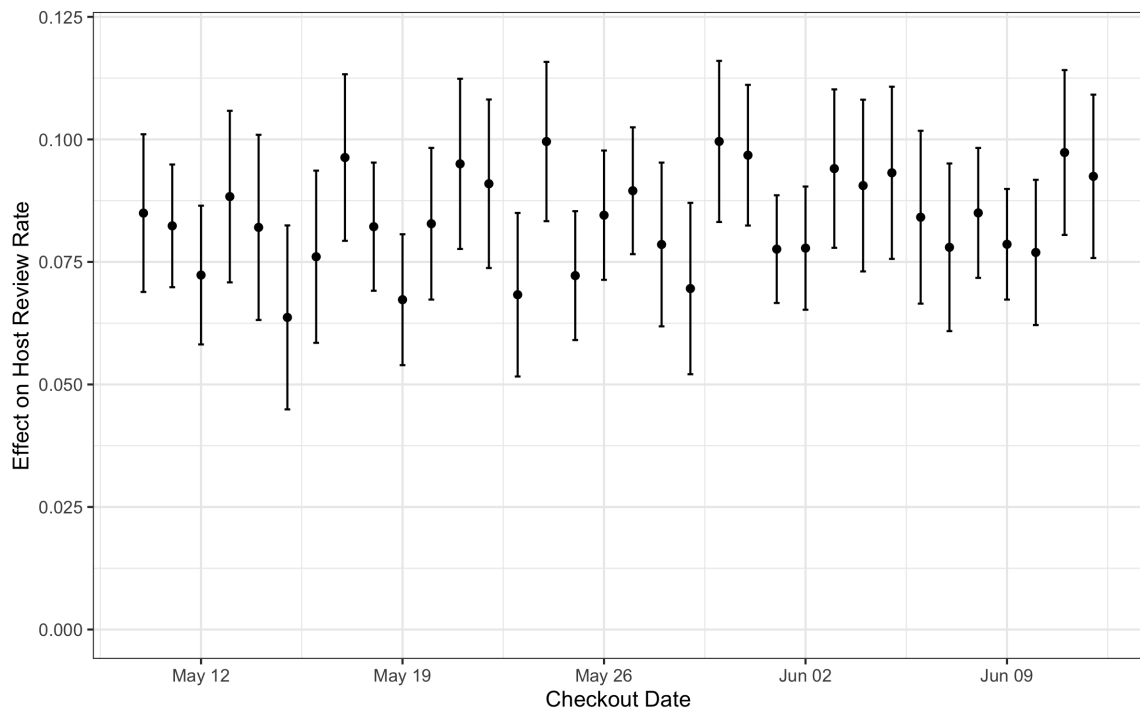
This figure plots the daily treatment effect on guest review rates and the 95% confidence interval. We use all transactions in the sample to increase precision.
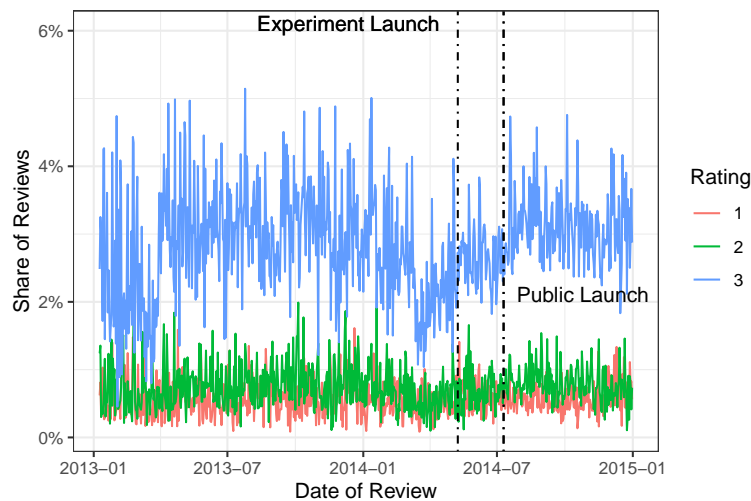
Figure A11: Review Rates Over Time
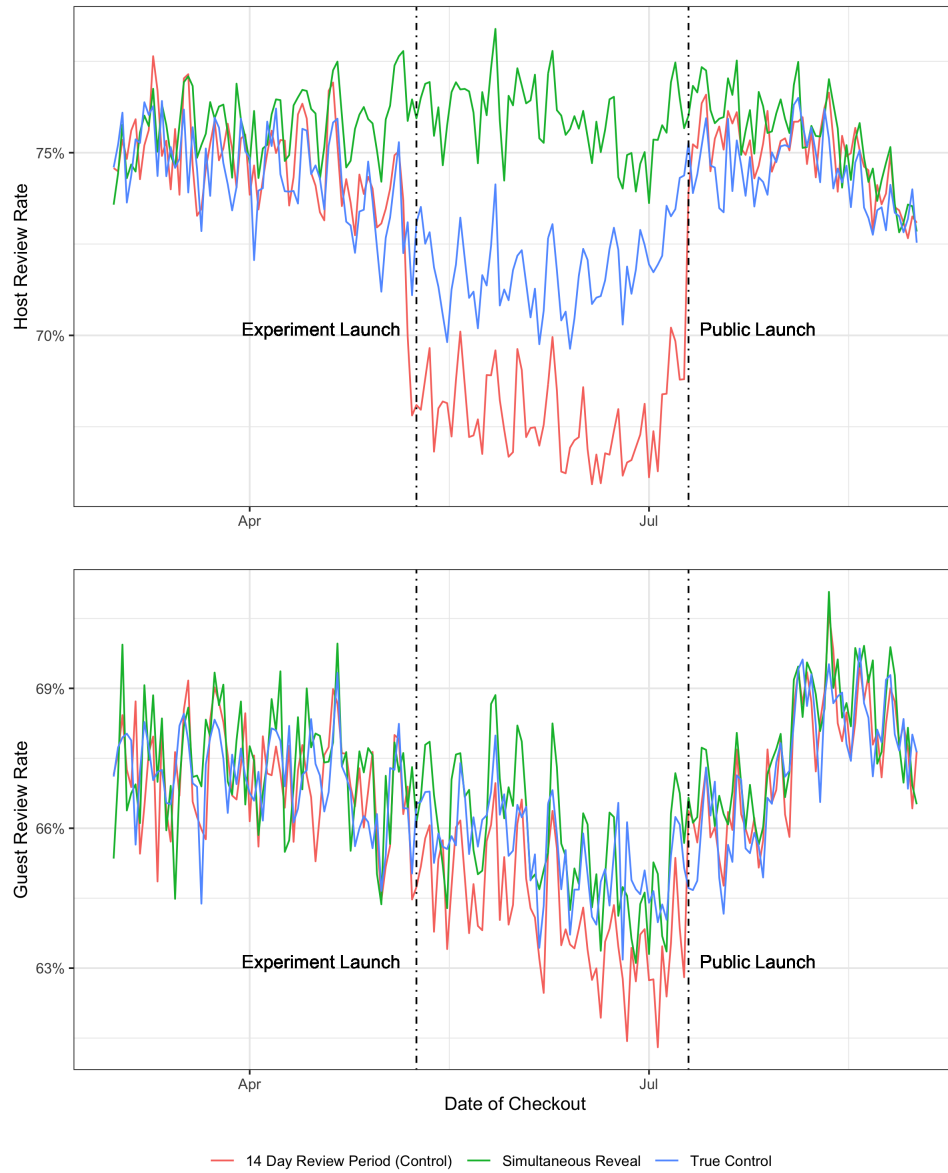
This figure displays the temporal trends of host and guest review rates over time by treatment group. Note that the Simultaneous Reveal Treatment changed the review period to 14 days from 31 days (True Control).

Figure A12: Ratings Over Time
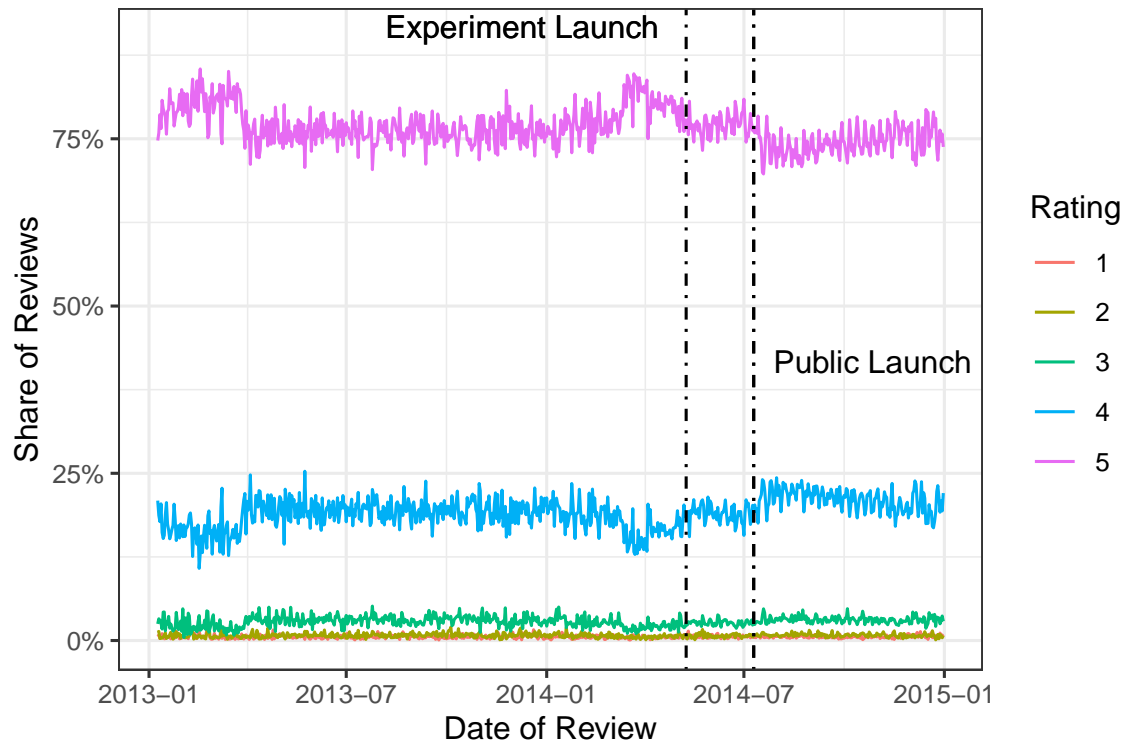
This figure displays the temporal trends of star ratings over time. Because the composition of guests and hosts varies with the growth of the platform, this figure includes only experienced guests reviewing from the domain ("www.airbnb.com") who stayed in a US based listing. The first vertical line demarcates 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 A13: 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.