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

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

[1]Boston University and MIT Initiative on the Digital Economy
[2]Airbnb, Inc.
[3]MIT Sloan School of Management

June 3, 2019

Abstract

Reciprocity in feedback may distort information and lead to inefficient outcomes in digital marketplaces. We analyze a large-scale field experiment on Airbnb in which the treatment was a simultaneous reveal review system that eliminated the ability of second reviewers to condition feedback on the content of the first review. The treatment had small negative effects on ratings and content, which points to a limited role of strategic reciprocity in shaping review content prior to the introduction of simultaneous reveal. The treatment also induced 9.7% more host reviews and 1.8% more guest reviews, consistent with a desire to reveal information about a counterparty's review by submitting a review. We fail to reject the null of no treatment effect on subsequent transactions by guests and hosts. Even in the absence of strategic reciprocity, there is evidence that some feedback may be distorted. For instance, when a guest does not recommend a host in a private and anonymous response, 19% of public reviews contain a high rating. Using plausibly exogenous variation in observational data, we show that this mismatch is partially explained by market participants interacting in a social manner and consequently omitting negative information in reviews.

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

[†]Primary Author: fradkin@bu.edu
[‡]Primary Experiment Designer: elena.grewal@airbnb.com
[§]dholtz@mit.edu

1 Introduction

Reputation systems are used by nearly every digital marketplace to reduce problems stemming from information asymmetry and moral hazard. They do so by soliciting information about transaction quality and displaying it to other market participants. However, the submission of accurate reviews by market participants is voluntary and costly, causing them to be under-provided in equilibrium (Avery, Resnick and Zeckhauser (1999); Miller, Resnick and Zeckhauser (2005)), with potential for non-random selection into who reviews. This leads to missing information and a variety of biases, which can affect outcomes for both buyers and sellers.[1] For instance, prior work shows that buyers who transact with sellers with biased reviews are less likely to transact on the platform in the future (Nosko and Tadelis (2015)). These factors make the design of effective reputation systems important for digital platforms.

We consider reputation system design in the setting of peer-to-peer marketplaces for services. This sector is often called "The Sharing Economy" and includes transactions for short-term rentals, ride-sharing, home improvement, and dog sitting. Transactions in this sector may be particularly susceptible to misreporting in reviews for at least two reasons. First, because the reputation systems in peer-to-peer marketplaces are bilateral, reviewers may choose to reciprocate or retaliate based on the content of the counterparty's review. This phenomenon, which we call strategic reciprocity, may cause reviewers to misreport their experiences both in an initial review and when retaliating.[2] Second, because buyers and sellers often interact in person, they may misreport their review out of a sense of social obligation. We call this phenomenon socially induced reciprocity.

We study both of these phenomena using experimental and observational analysis of reviewing behavior on Airbnb, a peer-to-peer marketplace for short-term lodging.[3] Our experimental treatment eliminated strategic reciprocity by preventing reviewers from seeing each other's reviews prior to writing their own. Treated hosts were assigned to a policy that hid reviews until either

[1] For a non-exhaustive list of documented biases in the literature, see these references: Bohnet and Frey (1999*a*); Cabral and Hortaçsu (2010); Dellarocas and Wood (2007); Horton (2014); Moe and Schweidel (2011); Nagle and Riedl (2014); Mayzlin, Dover and Chevalier (2014); Saeedi, Shen and Sundaresan (2015).

[2] There is evidence that strategic reviewing behavior may be important in the context of eBay (Bolton, Greiner and Ockenfels (2012); Cabral and Hortaçsu (2010); Saeedi, Shen and Sundaresan (2015)).

both guest and host had submitted a review or 14 days had elapsed. Another group of hosts were assigned a treatment where they had 14 days to review but the reviews were revealed to the counterparty as soon as they were submitted.[4] We treat this second group as the control group for the rest of our analysis.

The treatment increased the rate at which guests review by 1.2 percentage points (pp) and increased the total number of reviews written by guests with text labeled as negative by 1pp. The overall number of five star reviews did not change but the relative rate of five star reviews conditional on a review decreased by 1.2pp. There was also a 7pp increase in the rate at which guests left private suggestions to hosts. On the host side, the treatment increased review rates by 7pp but we estimate values close to 0 for the effect on reviews with negative text. In summary, both guests and hosts in the treatment are more likely to review and but the overall effect on the content of a review is small.

There are two mechanisms that combine to explain these results. First, the elimination of strategic reciprocity causally changes the content submitted by guests. We find a decrease of retaliatory reviews and an increase in private comments to a host after a transaction. The increased willingness of guests to submit private comments in the treatment suggests that guests in the control were worried about these comments potentially triggering retaliation. Second, the simultaneous review system increases the incentive for both parties to review each other. This occurs because submitting a review triggers the display of the other review on the website. To the extent that reviews matter and are mostly positive, having them visible on the platform earlier is advantageous. Curiosity may also play a role, since both guests and hosts would like to know what the counterparty's review says. We believe this explains the higher review rate and lower average time to review in the treatment.

Next, we document social reciprocity in reviews of hosts by guests. Stays on Airbnb frequently involve a social component, and internal Airbnb surveys of guests suggest that the social aspect

[3]For an overview of the economics of Airbnb see Farronato and Fradkin (2018).

[4]The experiment also had an arm in which users had 30 days to review and reviews were revealed as soon as submitted. This was the status quo policy on the platform. We don't focus on this group in the paper.

of Airbnb affects reviewing behavior. Work in behavioral economics also suggests that social distance and communication moderates giving behavior in giving games (Hoffman, McCabe and Smith (1996), Andreoni and Rao (2011), Bohnet and Frey (1999*b*)). A sense of social obligation can cause a reviewer to not want to hurt the reviewee's reputation. Social considerations may also cause reviewers to not submit negative reviews due to the potential for socially awkward interactions between reviewer and reviewee.

We cannot directly observe the amount of social interaction between guests and hosts. However, we do observe two proxies correlated with the degree of social interaction between guests and hosts. The first is whether a trip was to a private room within a home or to an entire home; stays in a private room are more likely to result in social interaction because of shared space. The second is if a host is a multi-listing host (defined as a host with more than 3 listings) or a casual host (those with 3 or fewer listings); multi-listing hosts are less likely to interact with guests because they typically do not reside in the property. We use plausibly exogenous variation in these proxies to measure socially induced reciprocity. Specifically, the same property can be rented as an entire home or as a private room and can be managed by a multi-listing host or a single-listing host. Individual properties exhibit transitions in this status over time, which allows us to identify the effects of these transitions. Social reciprocity should be bigger for transactions with a more social component, such as those in private rooms managed by single-listing hosts.

Prior work in behavioral economics also suggests that the degree of anonymity affects social giving behavior (Bohnet and Frey (1999*b*)). We hypothesize that socially induced reciprocity should have the greatest impact on review components most salient to the host because they are most likely to be attributed to the guest. The information most easily identified with the guest and most salient to the host is the review text, which is publicly visible next to a photo and a link to the guest's profile. Star ratings are somewhat salient to hosts, since they are displayed as rounded averages across all reviews on a listing's page. Hosts could potentially attribute a change in star ratings to a particular guest if the rounded rating changes but this is less likely than the attribution of text. Finally, guests submit information that is never publicly visible to anyone other than Airbnb

including customer service complaints and recommendations.

We measure the effects of social reciprocity by observing mismatch between signals of transaction quality. Suppose we have a proxy of low transaction quality such as a rating of less than 4 stars. Our theory of social reciprocity suggests that conditional on this low rating, transactions with more social reciprocity will result in more positive text reviews (which are more salient to hosts than star ratings). Consistent with this theory, we find that stays with multi-listing hosts at entire homes are 9.7pp more likely to have a negative text review than stays at casual hosts in private rooms. We corroborate this effect using the identification strategy described above, which uses host transitions to multi-listing status over time. We also document similar effects on mismatch between customer service complaints and star ratings, and between recommendations and star ratings. Our estimates are a lower bound on the true effect of socially induced reciprocity since stays with multi-listing hosts may exhibit social components and stays with other hosts may involve no social interaction.

Lastly, we consider the subsequent effects of the simultaneous reveal policy. Several papers have argued that reviews and ratings have powerful effects on market participants (e.g. Chevalier and Mayzlin (2006), Luca (2013), Pallais (2014)). We investigate whether the changes in reviews due to the simultaneous review system had effects on subsequent listing and guest outcomes such as bookings and nights. We cannot reject the null hypothesis of no effects for these outcomes and our 95% confidence intervals exclude large effects. There are three potential reasons for this lack of effect. First, as we've already argued, the simultaneous reveal system has small effects on star rating distributions. Second, the treatment helps some hosts, those who receive high ratings in the treatment but would not have in the control, but hurts other listings, who receive low ratings in the treatment but would not have in the control. Third, the simultaneous reveal policy was launched to the entire platform two months after the start of our experiment. If differences in outcomes between review systems take longer than 2 months to materialize, then we may be missing some of the effects of the simultaneous reveal policy.

The paper most closely related to ours is Bolton, Greiner and Ockenfels (2012), which exper-

imentally studies the effects of a simultaneous reveal system in the lab. The laboratory treatment greatly decreased the correlation between buyer and seller ratings and also reduced review rates. We conduct and analyze the first field experimental test of such a system and find small effects on ratings and *increases* in review rates. The differences in our results highlight the limitations of lab experiments, which may miss important features of the economic environment of a proposed policy (Levitt and List (2007)). Potentially important differences between the lab and our field setting include the social nature of the transaction, the underlying distribution of transaction quality, and the incentive to have reviews revealed quickly. Our experimental results also suggest that simultaneous reveal policies do not substantially affect reputation inflation (Horton (2014)). Observational studies comparing star ratings across accommodation websites (e.g. Zervas, Proserpio and Byers (2015)) have also led some to argue that Airbnb ratings are higher due to strategic considerations. We show that strategic considerations do not substantially affect reviewing on Airbnb, although socially induced reciprocity may have an effect.

Several papers have studied related aspects of reviews and trust in peer-to-peer markets. Klein, Lambertz and Stahl (2016) and Hui, Saeedi and Sundaresan (Forthcoming) study the effects of Ebay's change from a two-sided to a (mostly) one-sided reputation system on adverse selection and moral hazard, coming to different conclusions. We find that the treatment does not substantially affect host transaction volume, which may point to the importance of institutional context when studying these effects. Abrahao et al. (2017) studies the determinants of trust on Airbnb. An earlier, working version of this paper initially presented in 2014 contained many of the results presented in this work, and has influenced subsequent research regarding reputation on Airbnb including Proserpio, Xu and Zervas (2018) and Jaffe et al. (2017). Proserpio, Xu and Zervas (2018) propose that the social aspect of Airbnb transactions may affect realized quality in addition to reviewing behavior, while Jaffe et al. (2017) show how transactions with low quality sellers reduce guests' subsequent usage of the Airbnb platform. Previous work has also shown that reciprocity may exist in content platforms, and that bias can increase when transactions or tasks include a social element. 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) study observed reciprocity levels in content platforms relative to the null of no reciprocity and reject this null.

The rest of the paper proceeds as follows. In Section 2 we describe the setting for our work and in Section 3 we discuss the design of the simultaneous reveal experiment. In section 4 we conduct the experimental analysis and in Section 5 we document socially induced reciprocity. Section 6 presents results about the effects of the experiment on subsequent outcomes for hosts and guests and Section 7 discusses the implications of our findings for reputation system design.

2 Setting

Airbnb describes itself as a trusted community marketplace for people to list, discover, and book unique accommodations around the world. Airbnb has seen over 400 million guest arrivals since 2008 and contains over five million listings. 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 checked out of the listing, there was 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 could review each other.[5] Both the guest and host were prompted to review via e-mail the day after checkout. The host and guest also saw reminders to review their transaction partner if they log onto the Airbnb website or open the Airbnb app. Reminders were also sent when the counter-party left a review, or if the reviewer had not left a review after certain, pre-determined lengths of time.

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

[5]There are 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.

a one to five star rating,[6] 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 overall score and the sub-scores were displayed on each listing's page once there are at least three submitted reviews. The second page also contained an anonymous question that asked whether the guest would recommend staying in the listing being reviewed.

Overall ratings and review text were required and logged more than 99.9% of the time conditional on a guest review. Whether the other ratings are 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 6% of guest reviews. On Android, the sub-ratings were required but the anonymous recommendation was not logged. 79% of guest reviews were submitted via a desktop browser in our sample. We conduct most of our subsequent analysis in terms of binary variables (e.g., did this review have a 5 star overall rating or not?). Unless otherwise noted, we let missing values equal to 0 in these binary variables since this allows us to avoid conditioning on the device of the user.

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 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 were 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 to Airbnb.

[6] 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.

3 The Simultaneous Reveal Experiment

We now describe the design and effects of the simultaneous reveal experiment. Prior to May 8, 2014, both guests and hosts had 30 days after the checkout date to review each other and any submitted review was immediately posted to the website. This allowed for the possibility that the second reviewer retaliated against or reciprocated the first review. Furthermore, because of this possibility, first reviewers could strategically choose to not review or 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 were also 14 days to review. 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. In all cases, 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 receive different emails about reviews relative to users in the control. Figures 4 and 5 show the emails received by guests upon the end of the stay and when the counterparty has left a review first. Figure 6 shows the analogous first email for hosts.[7] Furthermore, both guests and hosts received a prominent notification before starting a review (Figure 3). There is a possibility that the copy of the email affected outcomes in a manner independent of the simultaneous reveal policy. While we can't address this possibility directly, we note that our prior is that details of the email had small effects. Therefore, it is unlikely that the simultaneous reveal policy had large effects that were canceled out by the change in email copy.

[7] Airbnb may have been changing some details of the email (such as the exact wording or images used in the email))

3.1 Description of Reviewing Behavior in the Experimental Sample

We now describe reviewing behavior in the 14 day treatment (henceforth called the control) and the simultaneous reveal treatment (henceforth called the treatment). To isolate just the effect of the simultaneous reveal experiment on reviews, we focus on the first transaction observed for each host either in the treatment or in the control. Guests and hosts in this sample do not know about the change to the review system before the trip, so changes we see in reviewing behavior are driven by the change in the review system rather than selection into who transacts with whom based on their assigned reputation system. Furthermore, this sample restriction allows us to avoid issues due to spillovers between multiple listings managed by the same host. We later turn to the effects of the experiment on subsequent stays and reviews.

Our baseline sample consists of 118,824 transactions starting with checkout dates on May 10, 2014 and ending with checkout dates on June 12, 2014. The sample excludes transactions in the experiment with checkout dates between June 6, 2014 and June 7, 2014 due to treatment logging issues.[8] Table 1 displays the mean outcomes in the treatment and control. Turning first to reviews in the control group, 67% 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 3.65 days to leave a review and guests taking an average of 4.25 days. Hosts may review at higher rates and more quickly for several reasons. First, because hosts receive inquiries from other guests, they check the Airbnb website more frequently than guests. Second, because hosts use the platform more frequently than guests and rely on Airbnb to earn money, they have more to gain than guests from inducing a positive guest review. Lastly, guests may be traveling after completing a stay and could have a harder time submitting a review.

over this time period. We were not able to recover the entire set of emails used by Airbnb during the time period and display emails which we believe to be representative. We solicited these emails from users who transacted during the time period of the experiment. Airbnb also inserted an additional piece of content in some of the initial emails sent to hosts (the exact time at which this began is unclear to us). This content concerns how reviews have changed (Figure A3). To see whether this made a difference, we computed daily treatment effects on reviews and found that the effect is similar over the time period during which the content may have been changing (Figure A4).

[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

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. Host reviews are even more positive, with 82% of host reviews containing all five star category ratings. Both guests and hosts also submit positive recommendations over 97% of the time conditional on an observed submission of a recommendation. These high rates of recommendations are notable due to the fact that the answers are anonymous and are never seen by anyone other than Airbnb.

Figure 3a shows the distribution of star ratings for submitted reviews both conditional and unconditional on a non-recommendation. The distribution of ratings for guests who do not recommend is lower than the distribution of ratings for those that do recommend. However, in over 20% of cases where the guest does not recommend the host, the guest submits a four or five star rating. Therefore, guests sometimes misrepresent the quality of their experiences in star ratings.[9] This misrepresentation can occur purposefully or because the guests do not understand the review prompt. Although we have no way to determine whether reviewing mistakes occur, the fact that fewer than 5% of reviewers recommend a listing when they submit a star rating lower than four suggests that guests typically understand the review prompt.

Text comprises another important part of the review which we incorporate into our analysis. We trained a logistic regression model on pre-experiment data to classify the sentiment of reviews and to determine the words and phrases associated with negative reviews. A discussion of the training procedure can be found in Appendix B. In Figure 3b we show the share of reviews with negative text conditional on the star rating. Over 90% of 1 and 2 star reviews are classified as negative and these reviews contain the most common negative phrases over 75% of the time. Three star reviews have text that is classified as negative over 75% of the time. Therefore, we find that guests who are willing to leave negative ratings are also typically willing to leave negative text. With regards to four star reviews, the results are mixed. Guests write negatively classified text 45% of the time.

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.

[9] It may also be the case that reviewers would not recommend a listing to a particular friend but still think that it deserves a high star rating. One reason this may occur is if the listing is cheap and a guest's friends are would not want to stay there.

Therefore, the review frequently does not contain information about why the guest left a four star rating.

Lastly, even when guests leave a five star rating, they leave negative text 13% of the time. This is potentially due to three reasons. First, even when the experience is not perfect, the listing may be worthy of a five star rating. Guests in that case may nonetheless explain any shortcomings of the listing in the review text. Second, our classifier has some measurement error and this may explain why some of these reviews were classified as negative. Third, reviewers may have accidentally clicked on the wrong star rating. Host reviews of guests are nearly always positive so we do not pursue a similar decomposition.

4 The Effects of the Experiment

Table 1 displays the treatment effects and standard errors next to the baseline. Both guests and hosts review more in the treatment and submit reviews faster. The experiment increased the share of transactions with a guest review by 1.2pp and the share of transactions with a host review by 7pp. These effects may be due to several mechanisms, which we discuss in subsection 4.1.

In addition to its effect on the rate and speed of reviews, the treatment also changes the content of reviews. Figure 7 displays the effects of the experiment on specific guest rating actions both conditional and unconditional on a review. The overall effect of the experiment on the types of reviews submitted is small as a percentage of all reviews submitted. Nonetheless, we detect statistically significant increases in 3 and 4 star reviews and a decreases in 1 star reviews. There is also an increase in the number of reviews submitted with negative text and more private feedback text to the host in the treatment.[10]

The observed increase in 3 and 4 star reviews suggests that guests are willing to leave more negative reviews. One explanation for this is that some guests are changing the content of their reviews to be more negative. Another is that guests with more negative experiences are more

[10]The experiment also increased the number of guest reviews submitted with a mobile device by 1.2pp (Table AI). This shift in devices is too small to explain changes in the recommendation rate between the treatment and control.

likely to leave a review under the treatment. The fact that the unconditional rate of 1 star reviews decreases in the treatment demonstrates that at least some of the change we observe is driven by guests changing how they review.[11] This decrease in 1 star reviews also suggests that the simultaneous reveal treatment decreases retaliatory behavior among guests. If retaliation is decreasing, then we should expect to see an especially large decline in 1 star ratings by guests in cases where hosts submit a negative review. Indeed, conditional on a host submitting negative text first, 1 star ratings by guests fall from 7% in the control to 2.2% in the treatment (see Figure A5).

Figure 8 displays the effects of the experiment on specific host feedback both conditional and unconditional on a review. In general, the overall effects are small and similar to those on guest rating actions, i.e., the treatment induces more 3 and 4 star reviews and fewer 1 star reviews. Furthermore, conditional on a first guest review with negative text, 1 star ratings fall from 4.9% in the control to 1.3% in the treatment (see Figure A6).

4.1 Investigation of Mechanisms

The above findings suggest that both guests and hosts strategically change their reviewing behavior due to the simultaneous reveal treatment. In this section, we attempt to distinguish between two types of mechanisms that drive these effects. The first set of mechanisms concerns the elimination in the treatment of strategic reciprocity. If the second reviewer cannot see the first review, then they cannot condition the review content on that first review. This should reduce the correlation between review content in the treatment. Furthermore, this may reduce the fear of retaliation in first reviews. An alternative mechanism present in the simultaneous reveal treatment is the desire to trigger the display of review information. Namely, users may want to submit their reviews quickly because they want the counterparty's review to appear on their profile as quickly as possible. Reviewers may also be curious to see what their counterparty wrote. We find that the desire to reveal information explains many of the observed treatment effects which cannot be explained by

[11] If we assume that the treatment weakly increases review rates for all guests, then the decrease in the quantity of 1 star reviews cannot be explained by a model that simply changes selection into who reviews.

a reduction in strategic reciprocity.

We begin by considering the timing with which reviews are submitted. Bolton, Greiner and Ockenfels (2012) argue that reciprocity and/or retaliation should create a correlation in the timing of feedback in bilateral reputation systems and that this correlation should be lower in the simultaneous reveal treatment. If this was the main mechanism driving reviews on Airbnb, we would expect the timing of reviews to be correlated in the control and less so in the treatment. Furthermore, if guests and hosts were fearful of retaliation, we would expect users in the control to leave negative first reviews towards the end of the review period, at which point there is less time for the counterparty to retaliate. If this is the case, when retaliation does occur, both reviews should occur towards the end of the review period. Consequently, if the simultaneous reveal treatment was only operating through a reduction in retaliation, then there should be an especially large decrease in observations towards the end of the review period.

Figure 9 displays a heatmap of the timing of guest and host feedback in the control group conditional on both parties submitting a review. Reviews are concentrated near the diagonal and there is also a concentration of observations in the upper right-hand corner, which corresponds to cases in which both reviews occurred toward the end of the review period. These stylized facts are consistent with the explanation of Bolton, Greiner and Ockenfels (2012) but also admit other explanations. The reviews at the diagonal may be due to the first review serving as a reminder to leave a review for the counterparty. The observations in the upper right-hand corner may be due to reminder e-mails sent toward the end of the review period.

While data from the control group is consistent with Bolton, Greiner and Ockenfels (2012)'s framework, a comparison of review timing in the treatment group and the control group is not. Figure 10 displays a heatmap of the change in the percentage of reviews occurring in each timing cell due to the treatment. There is an increase in the number of observations on or near the diagonal, and a decrease in the number of observations in the off-diagonal cells. The decrease in off-diagonal observations comes from a variety of timing bins and not just from the observations in the upper right. These results are contrary to simple stories regarding a reduction in reciprocity

and retaliation. Instead, the above pattern suggests that users had a desire to reveal information, which is why they submitted reviews earlier and sooner after the counterparty submitted a review.

Additional evidence that observed treatment effects are driven by the desire to reveal information comes from studying the rate at which guests and hosts review each other, and the order in which they do so. Figure 11 displays the impact of the treatment on the order in which guests and hosts review, conditional on at least one party reviewing. The treatment induces a 5.8pp decrease in the share of transactions for which guests review, but hosts do not, and a 4.5pp increase in the rate at which both parties review and guests review first. While we cannot conclusively attribute these effects to particular mechanisms, it is plausible that hosts are reviewing more often, especially after a guest has reviewed, because they want to "unlock" the contents of a guest's review.

Guests' and hosts' desire to reveal information makes it difficult to use review timing data alone to detect a reduction in retaliation and/or reciprocity. However, changes in the valence and sentiment of reviews provide evidence that the treatment did in fact have an impact on strategic reviewing behavior. Figure 12 displays the difference in pairwise text sentiment of reviews both conditional and unconditional on the transaction having both reviews. The treatment reduces cases where both reviews are positive and especially increases the number of cases where the guest submits a negative review while the host submits a positive review. Figure 13 displays a grid of heatmaps that show changes in the distribution of guest and host review text sentiment, conditional on both reviews being left in a particular timing bucket. Across almost all timing buckets, there are fewer cases where both guest and host reviews are positive, and more cases where the host review is positive and the guest review is negative. This is consistent with both a reduction in the fear of retaliation and a reduction in positive induced reviews. Evidence for retaliation is visible in the top left cell of Figure 13, where there is an 8pp decrease in mutually negative feedback when guests review late in the review period and hosts review subsequently.

In addition to the visual evidence discussed above, we consider regression based evidence of the effect of the treatment on mutually negative and positive feedback as measured by recommendations, ratings, and text. Table 2 shows the effect of the experiment on mutually negative

feedback. There is a negative treatment effect for each measure, and the treatment effects for star ratings and recommendations are statistically significant. This confirms that the treatment reduces retaliation, although the magnitude of that reduction is small. Table 3 shows the effect of the experiment on mutually positive feedback. Mutually positive recommendations and stars increase in the treatment and mutually positive text decreases. While the increase in mutually positive feedback for stars and recommendations does not suggest a decrease in reciprocity, it is not inconsistent with it. Given that the treatment increases the number of positive reviews left by both guests and hosts, the rate of mutually positive feedback could go up even if there is less reciprocity. In contrast, the reduction in mutually positive text is hard to justify by any mechanism other than the presence of induced reciprocity in the control group which was reduced in the treatment. Nonetheless, the small magnitude of this effect (-.009) does not point to a large role for this type of reciprocity in determining the pattern of reviews on the platform.

To summarize, we find increases in review rates and the speed of reviews due to the simultaneous reveal treatment. These reviews are more likely to be on the diagonal and are only slightly more likely to be negative than reviews in the control. We are able to detect evidence of reciprocity and retaliation by looking at the incidence of mutually positive and negative feedback, but find small effects. Consequently, we believe that the desire to reveal information is the primary causal mechanism of our results rather than a reduction in strategic reciprocity.[12]

5 Misreporting and Socially Induced Reciprocity

In Section 4, we showed that the simultaneous reveal treatment did not have a large effect on the reviewing behavior of Airbnb guests. Even in the treatment group, there remained discrepancies in the valance of different types of feedback within the same review. For instance, guests in the

[12]One additional possibility is that guests and hosts use the reputation system to communicate with each other about transaction quality. For example, they may communicate that they have 'no hard feelings' about something that happened during the stay. While this is possible, we think it is unlikely to explain our results. Most communication between guests and hosts happens through other mediums, such as the Airbnb's messaging feature or apps such as WhatsApp, neither of which depend on the review system. If guests and hosts are communicating about experience quality, we believe it is more likely that they are using these channels to do so, rather than the review system.

simultaneous reveal treatment who submitted a negative recommendation left four- or five-star ratings 18.6% of the time and had text classified as negative 19.2% of the time. Furthermore, guests who submitted 1 to 3 star ratings, submitted text classified as positive 22% of the time. In this section, we argue that there is another factor, socially induced reciprocity, which drives guests to misreport their experiences.

We define socially induced reciprocity as a pattern of reviewing behavior arising from feelings of social obligation among market participants. These perceived social obligations may cause reviewers to omit upsetting or harmful information about their counterparties in their reviews. It may also cause reviewers to omit information that they perceive could cause awkwardness in future social interactions. We hypothesize that guests and hosts develop these feeling when interacting with each other, either in person or through a messaging interface. There are a number of ways that guests and hosts can interact before, during, or after a stay on Airbnb. In most transactions, guests and hosts communicate through Airbnb's messaging platform to determine the availability of the listing and to coordinate check-in details. Guests and hosts also often socialize during the stay itself. When guests stay in a shared or private room, they are likely to interact with their host in the home's shared spaces. Even when guests rent an entire home, the host might show the guest around town or give local advice. Our hypothesis that the social aspect of Airbnb effects reviewing behavior is supported by internal Airbnb surveys of guests. For example, one guest said: "I liked the host so felt bad telling him more of the issues." Another guest wrote, "I often don't tell the host about bad experiences because I just don't want to hurt their feelings".

The degree of misreporting due to socially induced reciprocity should vary by the degree of social interaction between the guest and host. Although we do not observe social interaction itself, we observe two variables correlated with social interaction: whether the listing is managed by a multi-listing host (defined as a host with more than 3 listings) and whether a listing is a private room[13] or an entire home. Multi-listing hosts are less likely to interact with guests than casual hosts because they manage many properties and typically do not reside in the properties they manage. Stays in a private room are more likely to result in social interactions because guests and hosts

share a physical space during the stay.

Our method for measuring misreporting relies on observing a mismatch between different signals of transaction quality. Namely, consider two signals of transaction quality, A and B. Suppose that signal B is either more damaging to the host's future business or is more salient to the host. Then, if socially induced reciprocity was present, individuals should be less likely to report negative experiences through signal B than through signal A. We consider mismatch to occur when A indicates a negative experience and B indicates a positive experience. We formalize this intuition in a stylized model discussed in Appendix E. Our model shows how transactions with a more social component will be more likely to have high signals B conditional on a low signal A. In contrast, conditional on a high signal A, there will be little difference in signal B between more and less social transactions.

As an example, one signal of negative transaction quality is whether a guest contacts customer support during the trip. Customer service complaints are not displayed on the platform and are therefore unlikely to be noticed by the host or to hurt them. In contrast, the star rating is displayed as a rounded average on the listing page and may be salient to the host if the review changes the displayed average. Mismatch occurs in this case when there is customer service complaint but the guest leaves a 4 or 5 star rating. It is also true that these two signals of transaction quality don't measure exactly the same thing. For example, there will be cases where a call to customer service is unrelated to a particular listing, such as a bug in the app. Our empirical strategy would break down if the rate at which customers contacted customer service for reasons unrelated to the listing differed across our proxies for the degree of social interaction. While this could be possible for one pair of metrics, our subsequent results rely on two proxies for social interaction and four proxies of transaction quality and the results across these proxies are consistent with our hypothesis of socially induced reciprocity.

We first show that simple observational evidence suggests that misreporting varies by the degree of social interaction. Figure 14 plots three measures of review mismatch conditional on

[13]Listings where the guest shares a home with others during the trip, but does not have a private room, are called shared rooms. Fewer than 2% of the listings in our sample are shared rooms.

whether the stay was with a causal host in a private room or with a multi-listing host in an entire place. These figures show that guests are less likely to submit five star ratings conditional on a non-recommendations and customer service complaints when they stay with a multi-listing host. Guests who stayed with multi-listing hosts are also more likely to submit negative text conditional on low ratings.

Although the above differences are suggestive, they may not be causal. Reviews across listing types may differ for reasons other than the degree of social interaction. Different listing or host types may have different qualities or attract different types of guests. We use two forms of variation in the data to control for these factors. First, we use variation within listing to study the effect of staying with a multi-listing host on reviewing behavior. This variation exists because hosts who initially began as casual hosts can create additional listings over time to become multi-listing hosts. We condition on listing fixed effects, so that unobservable differences between listings such as location and property characteristics are held constant. Furthermore, we add guest fixed effects to allow for the possibility that different types of guests stay in different types of properties. Second, hosts sometimes rent out a property as both a private room and an entire home. Other than the size of the room, the price, and the degree of social interaction, there should be minimal differences in the quality of the two listings. We add address-specific fixed effects to isolate the effect of staying in a private room.

We estimate versions of the following regression:

$$\begin{aligned} y_{glt} =& \alpha_0 + \alpha_1 Nonsocial_{lt} + \alpha_3 Neg_{glt} + \beta_{NS} Nonsocial_{lt} * Neg_{glt} \\ &+ \beta' X_{gl} + \gamma_{gla} + \epsilon_{gl} \end{aligned} \tag{1}$$

where y_{glt} is a measure (rating or text) of a positive review by guest g for listing l at time t, EH_l is an indicator for whether the listing is an entire home, $Nonsocial_{lt}$ is an indicator for whether the host is of a less social type (either a multi-listing host or an entire home), Neg_{glt} is an indicator for a negative experience, X_{gl} are guest and trip characteristics, and γ_{gla} is a set of fixed effects

potentially including listing, guest, address, and market depending on the specification. If socially induced reciprocity is a factor in reviewing, then we would expect that β_{NS} is negative. This means that conditional on a signal of negative transaction quality, multi-listing hosts and those with entire homes get lower ratings and text than casual hosts and private rooms.

We begin by estimating models taking advantage of variation in the multi-listing host status of a given listing. Our sample consists of all transactions with a review to a private room or entire home with a check-in date between January 1, 2012 and April 1, 2014. We use this larger sample in order to obtain enough repeated observations for both listing and guest fixed effects.

Table 4 displays the results of regressions with multi-listing host as the measure of the social content of the stay for three types of mismatch. All of the regressions contain year-month fixed effects and controls for the prior reviews of a listing, the night of trip, number of guests, guest prior bookings, and price. Columns (1) and (2) show the results where 'non-recommend' is the less salient signal and a high star rating is the outcome. Column (2) adds both listing and guest fixed effects to control for unobserved differences in the quality of listings and the types of guests. In both regressions, the interaction between multi-listing host and a non-recommendation is negative, as would be true if there was socially induced reciprocity. Guests staying with a multi-listing host are 1.8pp less likely to leave a high rating in specification 2. Columns (3) and (4) contain similar regressions, where the less-salient signal is whether a guest complained to customer service. Once again, there is a negative interaction between multi-listing host and the signal of a negative experience. Lastly, columns (5) and (6) contain regressions where the outcome variable is whether a review has positive text and the less salient measure is whether a review had a low star rating. We find that conditional on having a low star rating, hosts are 4.1pp more likely to receive a review with negative text.[14] The non-interacted coefficient on 'multi-listing host' is close to 0 in specifications with listing fixed effects, consistent with our simple model in which we wouldn't expect mismatch differences when the less salient signal is positive.

[14]One concern with the above specification is that there other factors determining mismatch other than whether the trip was with a multi-listing host. In Table AIV we regress measures of mismatch directly on multi-listing host status as well as other controls and fixed effects. We estimate negative effects for all three outcomes, two of which are precisely estimated.

We also conduct a similar analysis using as our measure of a less-social interaction whether the stay was in an entire property. This type of analysis is made difficult due to the fact that entire homes and private rooms may differ due to their locations and hosts, and that this may affect reviewing behavior. To account for this, we estimate Equation 1 on a sample of listings for which hosts have multiple listing types at the same address and include address fixed effects.[15] Table 5 displays regressions for the three signals with entire home as the proxy for a less social stay. We find results similar to the ones for multi-listing hosts. For example, we find that conditional on a non-recommendation, stays at entire homes are 6.3pp less likely to have a high rating.

Both of our identification strategies show that socially induced reciprocity does affect guest reviewing behavior on Airbnb. Our results likely understate the true effect of socially induced reciprocity for several reasons. First, even transactions with multi-listing hosts at entire homes may involve meaningful social interactions between guests and hosts. Consequently, even in those cases, guests may omit negative information ratings. Second, recommendations and customer service complaints may also be affected by socially induced reciprocity. We discuss the implications of these results for reputation system design in Section 7.

6 Effects on Subsequent Outcomes

In the previous two sections, we've shown that reciprocity affects reviewing behavior on Airbnb. We now discuss the effects of reviewing behavior on market outcomes for hosts and guests. In our experiment, reviews and ratings left in the treatment are, on average, different than those left in the control arm. We might expect that being exposed to simultaneous reveal reviews earlier[16] has persistent effects.

There is no unambiguous theoretical prediction regarding the effects of our experiment on subsequent outcomes.[17] The treatment increases review rates and the speed of reviews. This means that other Airbnb users will have access to more information about trips to treated listings during

[15] We do not add guest fixed effects for this specification due to the relatively small number of observations.

[16] Recall that the simultaneous reveal was launched to all of Airbnb two months after the start of the experiment.

our experiment. We first consider potential effects on the hosts. The additional reviews induced by the treatment could have a positive effect. Positive reviews will arrive more quickly, giving hosts more time to capitalize on them. Hosts may also be able to learn from negative feedback and improve their listing. On the other hand, the additional reviews induced by the treatment could have a negative effect. In cases where these reviews are negative, guests may use to the information to avoid reviewed hosts. Hosts may also get discouraged by negative feedback and remove their listing from Airbnb. Whether the net effect is positive or negative is an empirical question.

In order to avoid spillovers between listings managed by the same host, we study these effects using the same sample of listings as in the primary analysis. We consider outcomes both during the experiment and after. During the experimental period, we measure the number of trips, the number of nights, the price per night among subsequent bookings in the experimental time period, and the revenue for each listing. Subsequent to the experiment, we measure bookings through the end of 2014 and whether the listing is active at the beginning of 2015.

Table 6 displays the treatment effects estimates for the above listing outcomes. We find effects of -1.3% on nights transacted during the experimental period and .8% on trips in the experimental period. These effects are only statistically significant at a 10% level. We do not find effects statistically different from 0 for revenue, price,[18] bookings by 2015, and whether a listing was active at the beginning of 2015. Our standard errors are small for these estimates, meaning that the overall effect is likely to be small. In Table 7, we show the results of these regressions with controls for the number of listing managed by a host, the number of reviews at the time of the first experimental trip, the price of the first transaction, the listing type, the log of listing bookings as of April, 2014, and the log of the nights of the first trip. We generally find effects that are smaller in magnitude and cannot be distinguished from 0. We conclude that the average effect of the treatment was close to 0 on listings' subsequent business outcomes.

[17] 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.

[18] Note that there are fewer observations for the price variable. This is due to the fact that we can't measure transaction prices for hosts who did not transact after the initial transaction in the experiment.

One may also be concerned that the small average treatment effects mask heterogeneity. For example, a marginal positive or negative review my have large effects for a subset of hosts. Figure 15 compares the empirical CDF of bookings by the end of 2014 for both the treatment and control groups. We conduct a Kolmogorov-Smirnoff test on the equality of these distributions and fail to reject the null (p-val = 0.6296). The theoretical effect of the simultaneous reveal treatment on subsequent guest outcomes is also ambiguous for reasons similar to the reasoning for listing business outcomes. The analysis is slightly complicated by the fact that the treatment was assigned at a host level. In practice, this is not a problem since we can isolate a group of guests who were close to randomly assigned. These are guests whose first trip in the experiment was also the first trip for the host in the experiment. We run a regression analogous to the one in Table 6 for guests. Table 8 displays the results. We find a 1.2% decrease in nights by 2015 for guests in the treatment which is significant at a 10% level. The estimated effects for the other guest outcomes (nights in the experiment, trips in the experiment, and bookings by 2015) are smaller in magnitude and not statistically significant. Figure Figure 15 compares the empirical CDF of total guest bookings by the end of 2014 for both the treatment and control groups. We conduct a Kolmogorov-Smirnoff test on the equality of these distributions and fail to reject the null (p-val = 0.5599). Overall, we find no evidence that the simultaneous reveal experiment had a substantial effect on subsequent listing or guest market outcomes. This lack of effect may have several explanations. First, the effect of the simultaneous reveal treatment on the number and distribution of ratings was small. Consequently, there is unlikely to be a large aggregate effect on outcomes. Second, the effect of the treatment is theoretically ambiguous. Lastly, our analysis only allows us to measure the effect of an additional two months of exposure to the simultaneous reveal treatment. This effect is distinct from, and likely smaller than, the effect of a permanent difference in review policies.

7 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 within the review system. In this paper, we use an experiment and proprietary data from Airbnb to document reciprocity, its causes, and its role in determining the ratings distribution.

Our first set of results pertain to reciprocity caused by interactions within the review system, which we call strategic reciprocity. Namely, Airbnb began with a review system in which reviews were displayed immediately after submission. This review system allowed for reciprocity in response to positive reviews and retaliation against negative reviews. We study the simultaneous reveal intervention, in which reviews were hidden until both parties submitted a review or until 14 days had passed following the trip. This policy change precluded reciprocity and retaliation related to review content. We find that the simultaneous reveal review policy increased review rates and made the average review more negative, although the effects on review valence were small. Furthermore, the experiment had almost no effects on subsequent outcomes (bookings and nights) of market participants. We conclude that these strategic factors were not the primary reason for the observed distribution of ratings and reciprocity on the platform. Instead, we think that the primary mechanism by which the policy affected reviews was by creating an incentive for users to review faster and more frequently. This incentive was generated by the fact that a counterparty's review was hidden until both parties reviewed. This inducement of more reviews was likely to be good for the platform.

We also investigated another form of reciprocity related to social interactions between guests and hosts, which we call socially induced reciprocity. More specifically, social interactions before, during, or after a stay on Airbnb may lead market participants to omit relevant information from their reviews. We do find evidence that socially induced reciprocity affects ratings. This finding

suggests that a platform can improve the reputation system through designs which make reviewing less personal. For example, appeals to review based on the welfare of other future guests to a property may induce the guest to submit more informative ratings.

Our research points to several directions for future investigation. The high rates of five star ratings and recommendations suggest that most experiences on Airbnb are positive. Indeed, it would be difficult for the platform to be so successful if it didn't generate value for its users. This fact could motivate a research agenda that attempts to identify the market design features of Airbnb that facilitate high quality matches. For instance, Airbnb's trust and safety team may proactively filter bad actors from the market. There are several determinants of Airbnb's ratings distribution that we do not study. For instance, not all users submit reviews on Airbnb. If those that opt out of reviewing have lower quality experiences, reviews on the platform will tend to be more positive. 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. It would be interesting to explore mechanisms that allow reviewers to opt out of associating their review with their profile.

Finally, reviews may describe how an experience compared to the reviewer's own expectations, rather than describing an experience's absolute quality. For example, if a listing on Airbnb was cheaper than a particular hotel, then guests may not expect hotel quality amenities and service from the host. It may be possible to design review systems that separate out 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

Abrahao, Bruno, Paolo Parigi, Alok Gupta, and Karen S. Cook. 2017. "Reputation Offsets Trust Judgments Based on Social Biases among Airbnb Users." *Proceedings of the National Academy of Sciences*, 114(37): 9848–9853.

Andreoni, James, and Justin M. Rao. 2011. "The Power of Asking: How Communication Affects Selfishness, Empathy, and Altruism." *Journal of Public Economics*, 95(7-8): 513–520.

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

Bohnet, Iris, and Bruno S Frey. 1999*a*. "Social Distance and Other-Regarding Behavior in Dictator Games: Comment." *American Economic Review*, 89(1): 335–339.

Bohnet, Iris, and Bruno S Frey. 1999*b*. "The Sound of Silence in Prisoner's Dilemma and Dictator Games." *Journal of Economic Behavior & Organization*, 38(1): 43–57.

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

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

Chevalier, Judith A., and Dina Mayzlin. 2006. "The Effect of Word of Mouth on Sales: Online Book Reviews." *Journal of Marketing Research*, 43(3): 345–354.

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.

Farronato, Chiara, and Andrey Fradkin. 2018. "The Welfare Effects of Peer Entry in the Accommodations Market: The Case of Airbnb." *NBER Working Paper*.

Hoffman, Elizabeth, Kevin McCabe, and Vernon L. Smith. 1996. "Social Distance and Other-Regarding Behavior in Dictator Games." *American Economic Review*, 86(3): 653–60.

Horton, John J. 2014. "Reputation Inflation in Online Markets."

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

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

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

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.

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.

Luca, Michael. 2013. "Reviews, Reputation, and Revenue: The Case of Yelp.Com." *HBS Working Knowledge*.

Mayzlin, Dina, Yaniv Dover, and Judith Chevalier. 2014. "Promotional Reviews: An Empirical Investigation of Online Review Manipulation." *American Economic Review*, 104(8): 2421–2455.

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

Moe, Wendy W., and David A. Schweidel. 2011. "Online Product Opinions: Incidence, Evaluation, and Evolution." *Marketing Science*, 31(3): 372–386.

Nagle, Frank, and Christoph Riedl. 2014. "Online Word of Mouth and Product Quality Disagreement." Social Science Research Network SSRN Scholarly Paper ID 2259055, Rochester, NY.

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

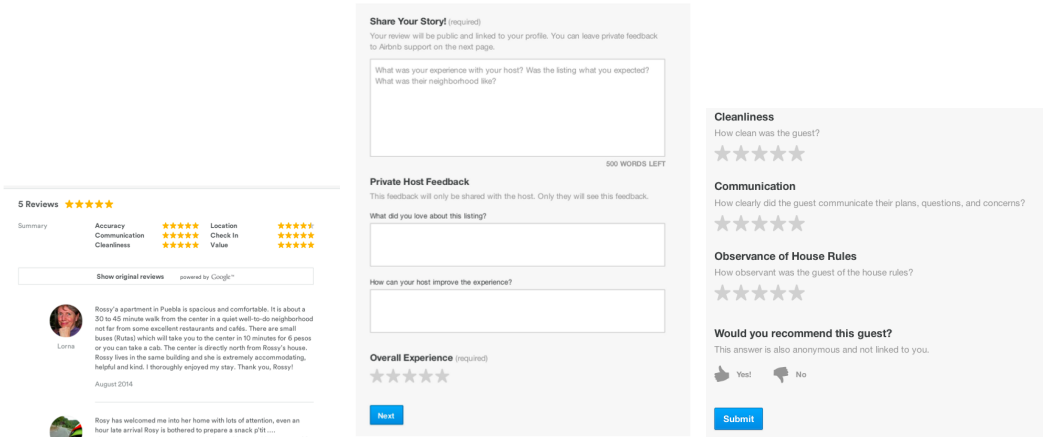
Pallais, Amanda. 2014. "Inefficient Hiring in Entry-Level Labor Markets." *American Economic Review*, 104(11): 3565–99.

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

Saeedi, Maryam, Zequian Shen, and Neel Sundaresan. 2015. "The Value of Feedback: An Analysis of Reputation System."

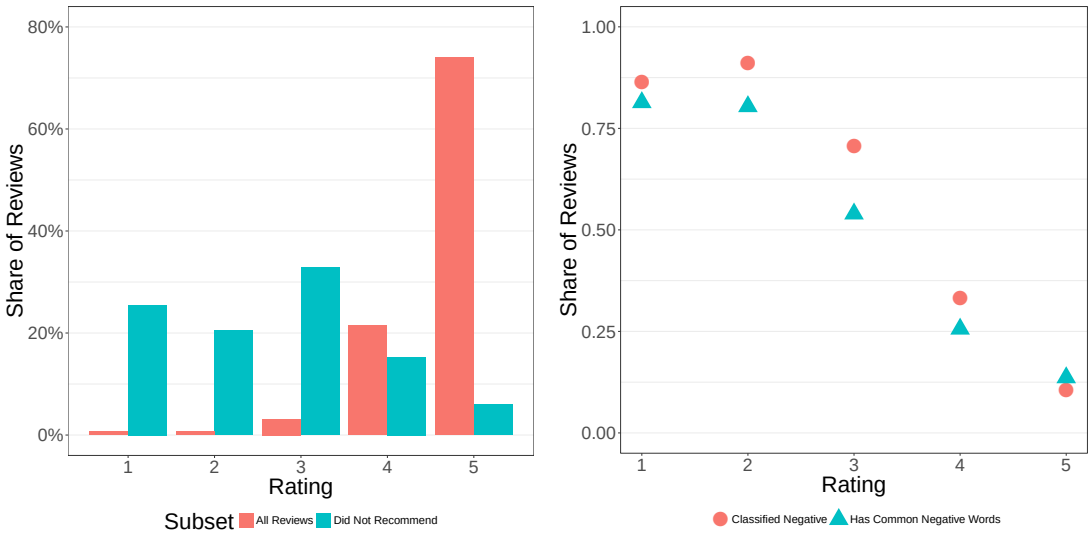
Zervas, Georgios, Davide Proserpio, and John Byers. 2015. "A First Look at Online Reputation on Airbnb, Where Every Stay Is Above Average." Social Science Research Network SSRN Scholarly Paper ID 2554500, Rochester, NY.

8 Figures



(a) Reviews on Listing Page (b) Review of Listing (Page 1) (c) Review of Guest (Page 2)

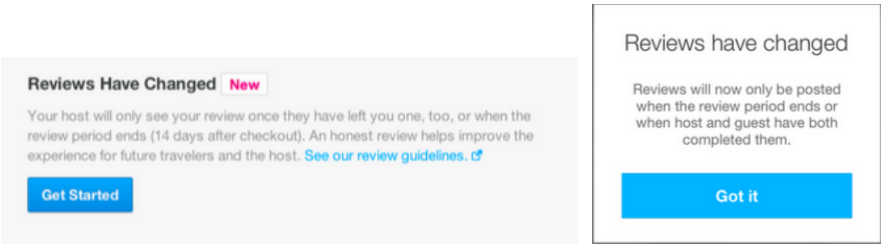
Figure 1: Review flow on the website



(a) Ratings Distribution (b) Text Conditional on Rating and Non-recommend

Figure 2: Ratings and Text Distributions

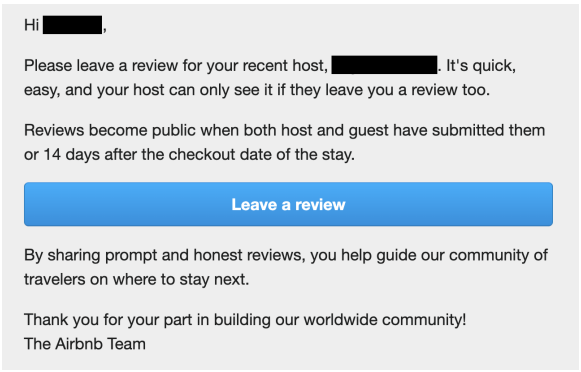
The left figure displays the distribution of submitted ratings in the control group of the simultaneous reveal experiment. Only first stays during the experimental period for each listing are included. The right figure displays the prevalence of negative text conditional on rating. "Classified Negative" refers to the classification by the regularized logistic regression based on the textual features of a review.



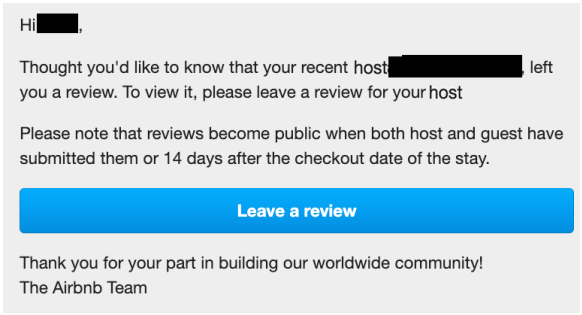
(a) Desktop (b) Mobile

Figure 3: Simultaneous Reveal Notification

The above figures display the notifications shown to guests. The desktop notification had the word 'host' replaced with the word 'guest'.

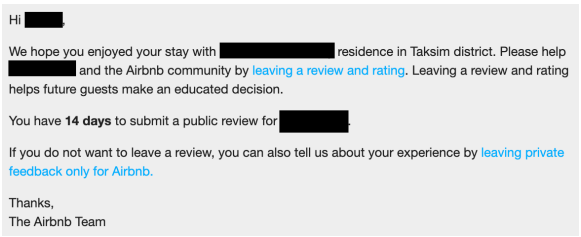


(a) First Email

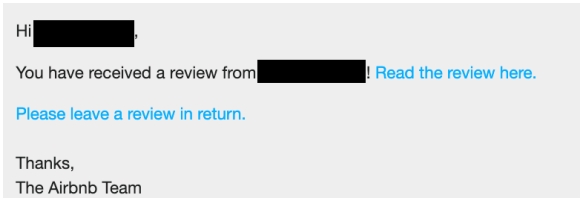


(b) Second Email

Figure 4: Simultaneous Review Email - Guest

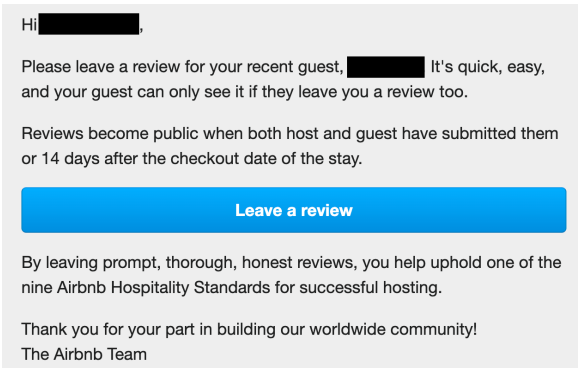


(a) First Email

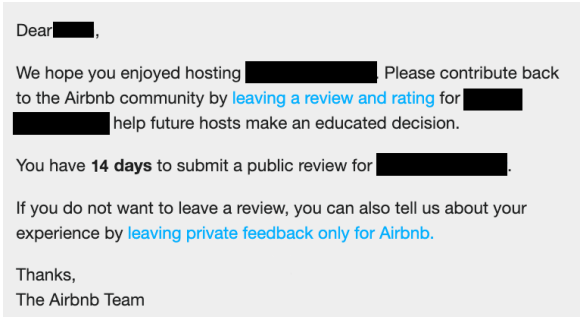


(b) Second Email

Figure 5: Control Email - Guest



(a) Treatment



(b) Control

Figure 6: Host First Emails

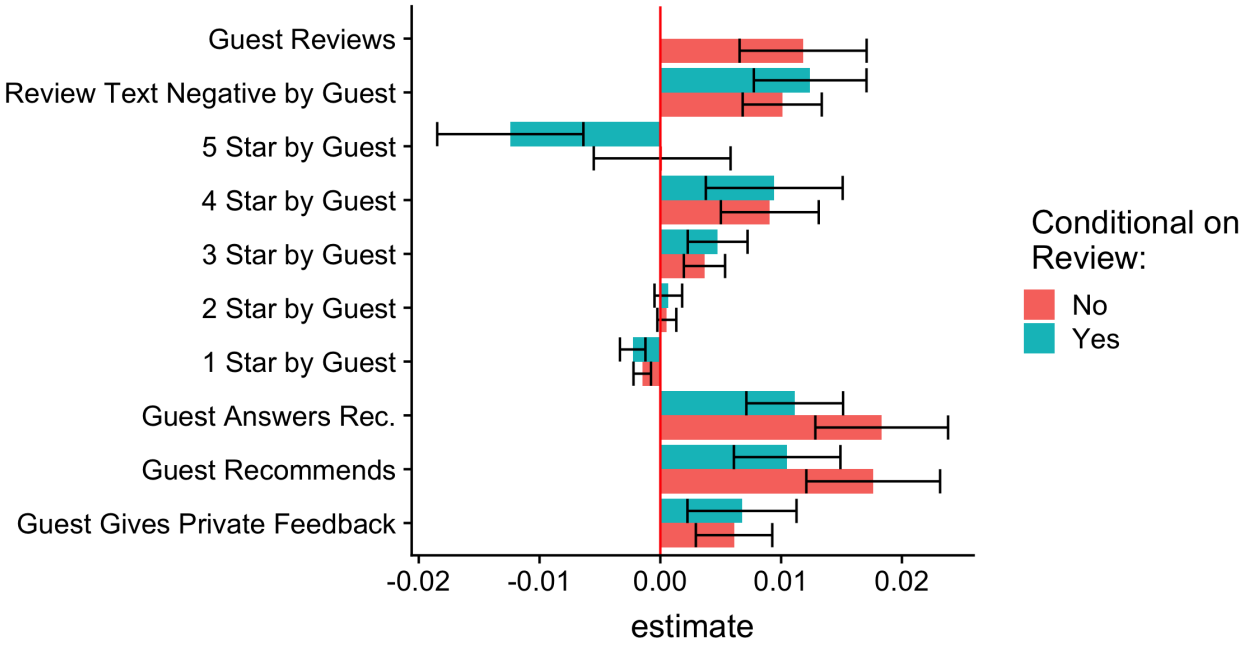


Figure 7: Effects of Experiment on Guest Reviews

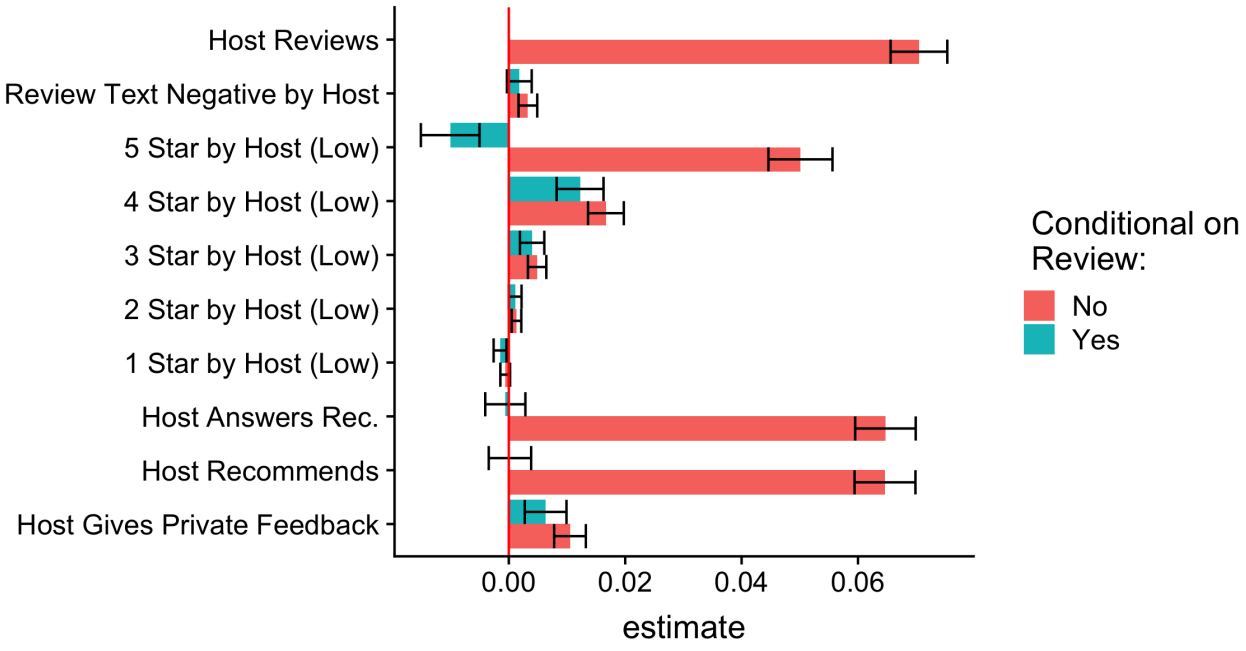


Figure 8: Effects of Experiment on Host Reviews

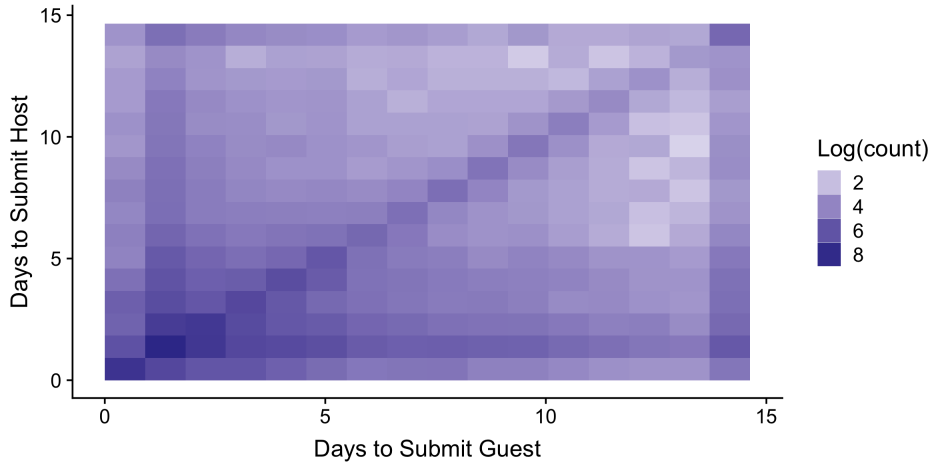


Figure 9: Ratings Timing

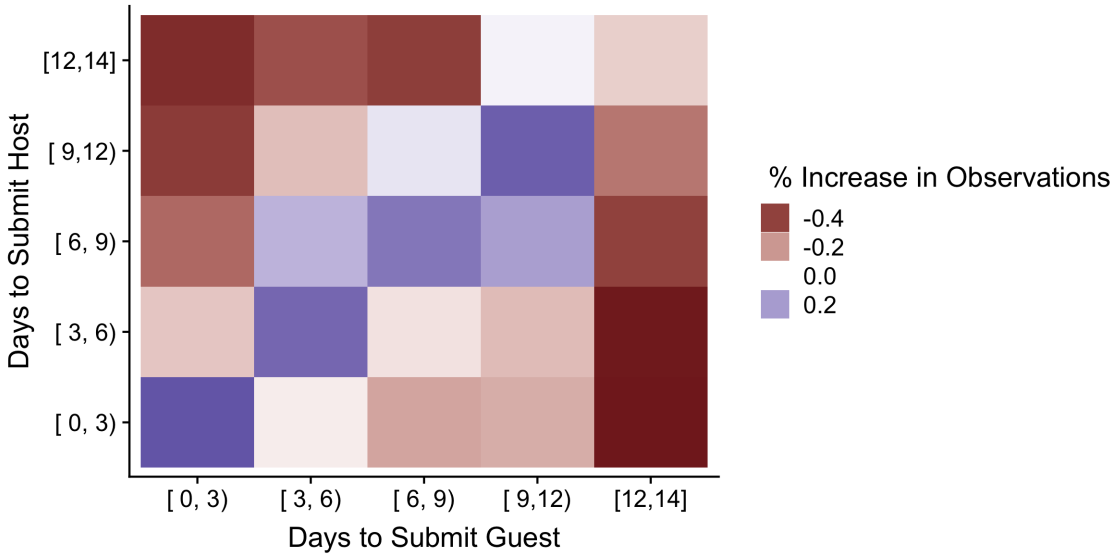


Figure 10: Changes in the Timing of Reviews

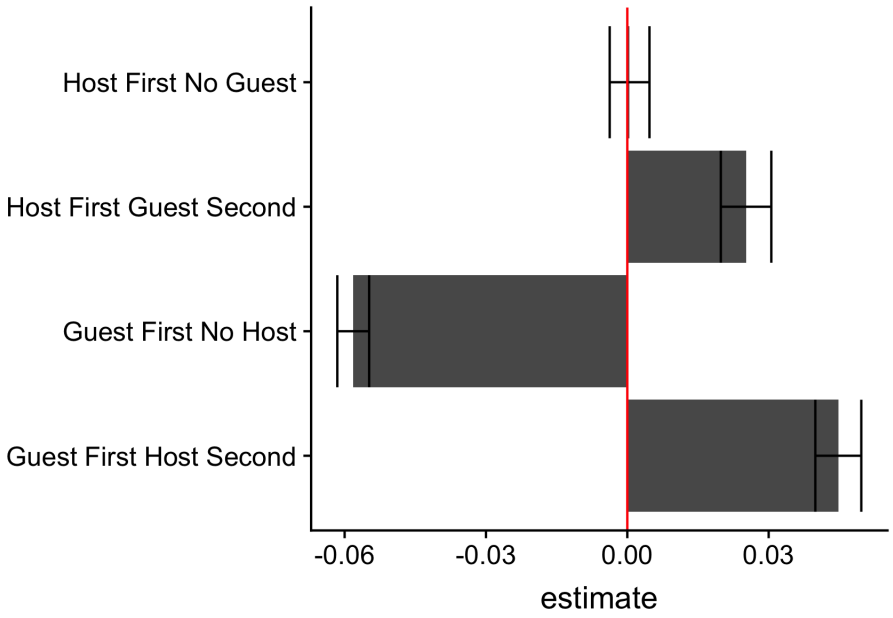


Figure 11: Timing Causal Effects

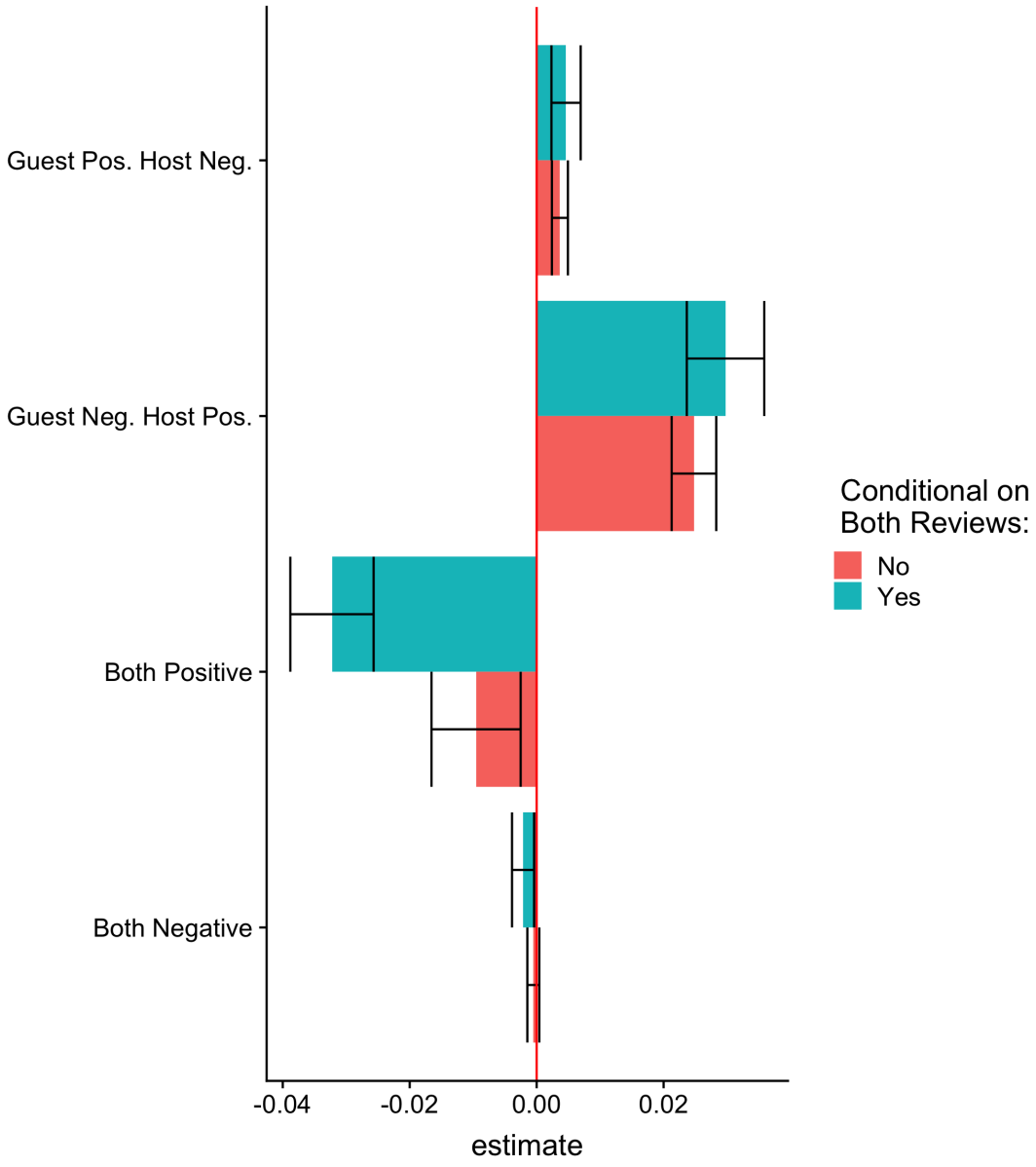


Figure 12: Changes in the Incidence of Pairwise Text Sentiment

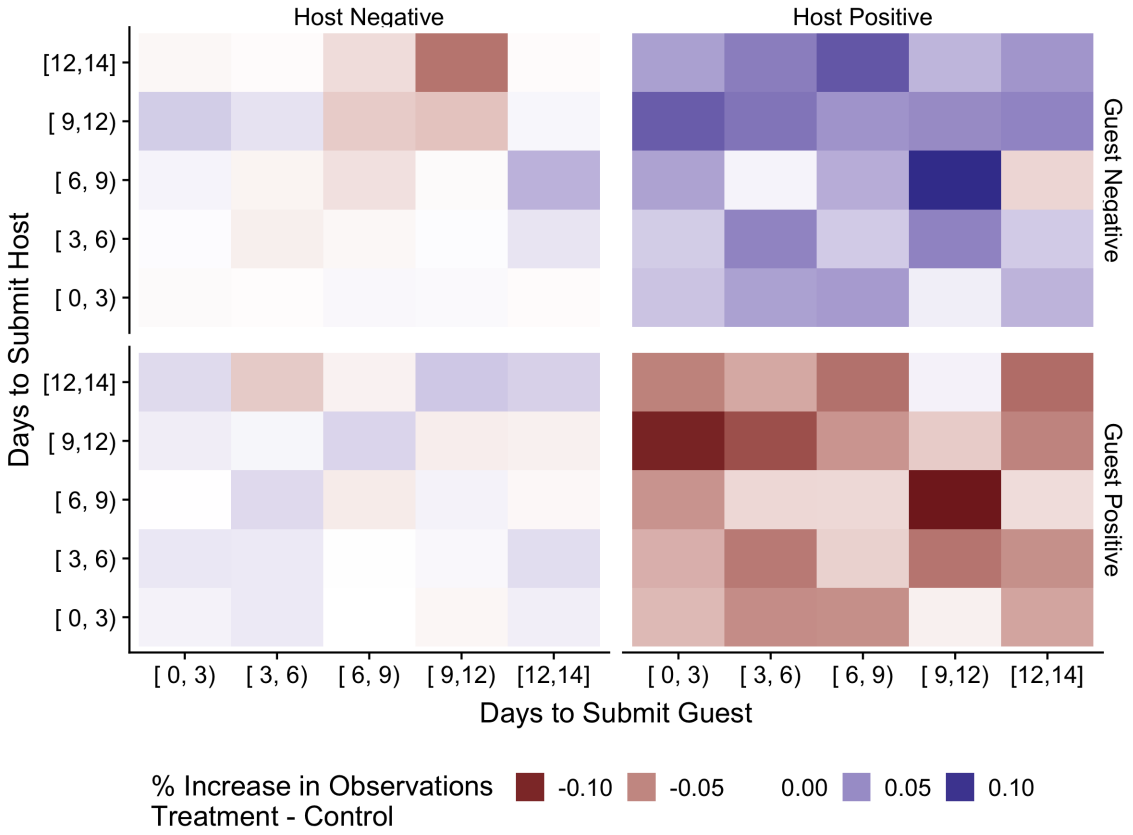


Figure 13: Changes in the Timing of Feedback Between Treatment and Control

Positive and negative reviews in the above figure are calculated in terms of the valence of text submitted in the review. To calculate the value of each cells, we take the difference in the share of observations in each cell between the treatment and control group.

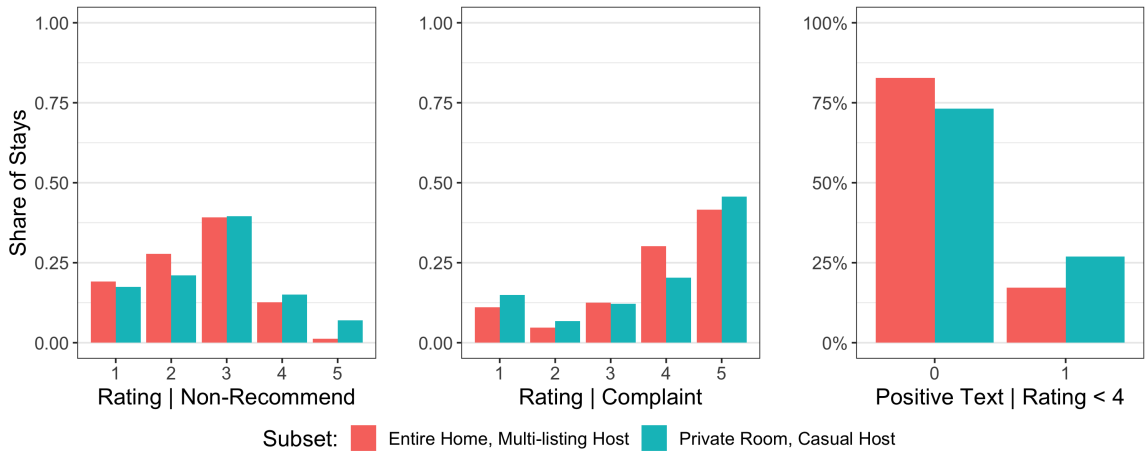
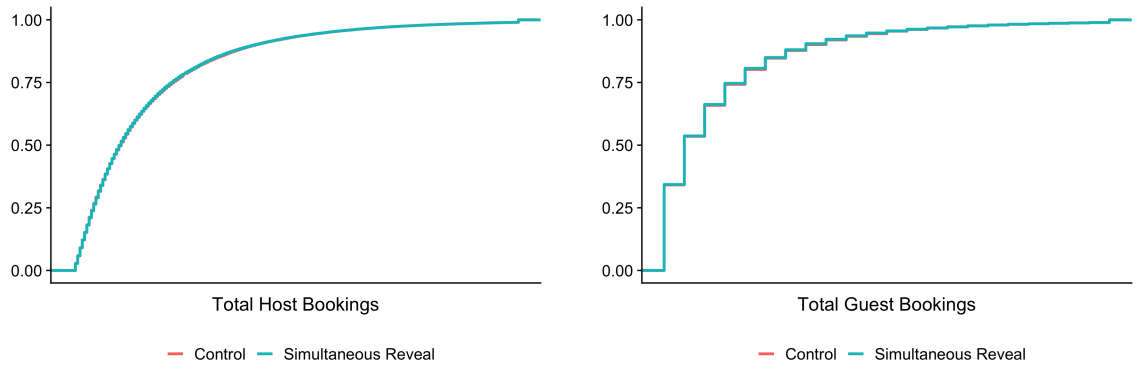


Figure 14: Conditional Ratings Differences

The above figure displays outcomes in the treatment group of the simultaneous reveal experiment. The figures plot, in order, the distribution of ratings conditional on a non-recommendation, the distribution of ratings conditional on a customer service complaint, and the probability of positive text conditional on a rating lower than 4 stars. The two comparison groups consist of stays in entire homes of multi-listing hosts (those who have more than 3 listings) and stays in private rooms of casual hosts (those who have fewer than 3 listings).



(a) Listing

(b) Guest

Figure 15: Distribution of Bookings by January 1, 2015

We censor the number of bookings at the 99^{th} percentile to make the figure easier to read.

9 Tables

	Host						**Guest**					
		All			*Conditional*			*All*			*Conditional*	
	Mean	Effect	SE	Mean	Effect	SE	Mean	Effect	SE	Mean	Effect	SE
Reviews	0.717	**0.070**	0.002				0.680	**0.012**	0.003			
Time to Review (Days)				3.831	**-0.411**	0.025				4.890	**-1.002**	0.039
Recommends	0.656	**0.065**	0.003	0.915	0.000	0.002	0.596	**0.018**	0.003	0.877	**0.010**	0.002
Overall Rating = 5							0.503	0.000	0.003	0.739	**-0.012**	0.003
Lowest Subrating = 5	0.591	**0.050**	0.003	0.824	**-0.010**	0.003	0.323	-0.001	0.003	0.476	**-0.010**	0.003
Review Text Negative	0.019	**0.003**	0.001	0.026	0.002	0.001	0.087	**0.010**	0.002	0.128	**0.012**	0.002

Table 1: Summary Statistics

This table displays the mean outcomes in the control groups (mean), the difference between the outcome in the treatment and control (effect), and the standard error of the effect estimate. The results are displayed for both hosts and guests and unconditional and conditional on a review. Values in bold are statistically significant at a 5% level.

	Mutual Negative Rec.	Mutual Negative Stars	Mutual Negative Text
	(1)	(2)	(3)
Treatment	-0.001^{*}	-0.001^{***}	-0.0005
	(0.001)	(0.0004)	(0.0004)
Dependent Variable Mean	0.008	0.005	0.004
Trip, Guest, and Host Characteristics	YES	YES	YES
Data and Market FE	YES	YES	YES
Observations	119,441	119,441	119,441
R^2	0.014	0.011	0.012

Table 2: Effects on Mutually Negative Feedback

This table displays the results of regressions of mutually negative review outcomes on the treatment and covariates. Covariates include the share of prior listing trips with a five star review, the number of guests, the number of nights, whether the guest was from the US, whether the guest had prior bookings, the listing types, whether the listing was a multi-listing host, the log of the number of reviews and bookings of the listing, the log of the number of bookings by the guest, and the log of the price per night.

	Mutual Positive Rec.	Mutual Positive Stars	Mutual Positive Text
	(1)	(2)	(3)
Treatment	0.063***	0.059***	−0.009***
	(0.003)	(0.003)	(0.003)
Dependent Variable Mean	0.482	0.527	0.44
Trip, Guest, and Host Characteristics	YES	YES	YES
Data and Market FE	YES	YES	YES
Observations	119,441	119,441	119,441
R^2	0.045	0.051	0.114

Table 3: Effects on Mutually Positive Feedback

This table displays the results of regressions of mutually positive review outcomes on the treatment and covariates. Covariates include the share of prior listing trips with a five star review, the number of guests, the number of nights, whether the guest was from the US, whether the guest had prior bookings, the listing types, whether the listing was a multi-listing host, the log of the number of reviews and bookings of the listing, the log of the number of bookings by the guest, and the log of the price per night.

	Rating > 3		Rating > 3		Positive Text	
	(1)	(2)	(3)	(4)	(5)	(6)
Multi-listing Host	−0.026***	0.001	−0.045***	0.0003	−0.081***	−0.004**
	(0.0003)	(0.001)	(0.0003)	(0.001)	(0.001)	(0.002)
Non-Recommend	−0.763***	−0.666***				
	(0.001)	(0.002)				
Customer Service			−0.096***	−0.069***		
			(0.001)	(0.001)		
Star < 4					−0.369***	−0.276***
					(0.001)	(0.001)
Multi-listing * Non-Recommend	−0.013***	−0.018***				
	(0.002)	(0.003)				
Multi-listing * Customer Service			−0.051***	−0.040***		
			(0.002)	(0.003)		
Multi-listing * Star < 4					−0.059***	−0.041***
					(0.001)	(0.002)
Trip, Guest, and Host Characteristics	Yes	Yes	Yes	Yes	Yes	Yes
Year-Month and Market FE	Yes	Yes	Yes	Yes	Yes	Yes
Listing and Guest FE	No	Yes	No	Yes	No	Yes
Obs	6923684	4146441	7933097	4937141	6430878	4011924

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

Table 4: Socially Induced Reciprocity - Multi-listing Hosts

The outcomes in the above regression are whether the guest's star rating is greater than 3 and whether the text of a review was classified as positive. The estimation is done on all trips with reviews between 2012 and 2014 for a 50% sample of guests. An additional criterion was that the respective mismatch variables were non-missing. Additional controls for the number of prior reviews, nights of trip, guests in trip, number of prior bookings by the guest, and price per night are included. The number of observations decreases in columns (2), (4), and (6) due to the fixed effects included in the specification. *p<0.10, ** p<0.05, *** p<0.01

	Rating > 3				Positive Text	
	(1)	(2)	(3)	(4)	(5)	(6)
Entire Home	-0.008^{***}	0.003	-0.018^{***}	-0.00000	-0.027^{***}	0.013^{**}
	(0.002)	(0.003)	(0.003)	(0.003)	(0.005)	(0.005)
Non-Recommendation	-0.727^{***}	-0.688^{***}				
	(0.010)	(0.010)				
Customer Service			-0.089^{***}	-0.077^{***}		
			(0.008)	(0.008)		
Rating < 4					-0.324^{***}	-0.276^{***}
					(0.007)	(0.007)
Entire Home * Non-Recommendation	-0.062^{***}	-0.063^{***}				
	(0.012)	(0.014)				
Entire Home * Customer Service			-0.036^{***}	-0.035^{***}		
			(0.011)	(0.011)		
Entire Home * Rating < 4					-0.054^{***}	-0.043^{***}
					(0.010)	(0.010)
Trip, Guest, and Host Characteristics	Yes	Yes	Yes	Yes	Yes	Yes
Year-Month and Market FE	Yes	Yes	Yes	Yes	Yes	Yes
Address FE	No	Yes	No	Yes	No	Yes
Observations	133,754	133,754	152,598	152,598	129,452	129,452
Note:						*p<0.1; **p<0.05; ***p<0.01

Table 5: Socially Induced Reciprocity - Entire Home, Address Fixed Effects

The estimation for this regression is done for trips by guests to hosts who had at least one private room and one entire home listing at the same address. Additional controls for the number of prior reviews, nights of trip, guests in trip, number of prior bookings by the guest, and price per night are included. *p<0.10, ** p<0.05, *** p<0.01

	Dependent variable:					
	Log(Nights in Exp.)	Log(Trips in Exp.)	Log(Rev. in Exp.)	Log(Avg. Price in Exp.)	Bookings by 2015	Active in 2015
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment	-0.013^{*}	-0.008^{*}	-0.035	-0.006	-0.001	-0.003
	(0.007)	(0.004)	(0.021)	(0.006)	(0.006)	(0.003)
Constant	1.331^{***}	0.781^{***}	4.361^{***}	5.111^{***}	3.030^{***}	0.746^{***}
	(0.005)	(0.003)	(0.015)	(0.004)	(0.004)	(0.002)
Observations	119,550	119,550	119,550	73,234	119,550	119,550
R^2	0.00003	0.00003	0.00002	0.00001	0.00000	0.00001
Note:						*p<0.1; **p<0.05; ***p<0.01

Table 6: Long-term Listing Outcomes

This table plots the effect of the experiment on listing outcomes. Only the first listing booked in the experimental period per host is included.

	Dependent variable:					
	Log(Nights in Exp.)	Log(Trips in Exp.)	Log(Rev. in Exp.)	Log(Avg. Price in Exp.)	Log(Bookings by 2015)	Active in 2015
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment	−0.008	−0.005	−0.023	−0.006	0.003	−0.003
	(0.006)	(0.004)	(0.018)	(0.004)	(0.003)	(0.002)
Observations	119,550	119,550	119,550	73,234	119,550	119,550
R^2	0.296	0.321	0.246	0.459	0.688	0.079
Note:						*p<0.1; **p<0.05; ***p<0.01

Table 7: Long-term Listing Outcomes - Controls Added

This table plots the effect of the experiment on listing outcomes. Only the first listing booked in the experimental period per host is included. Covariates include the log of prior reviews, whether the listing had a prior review, the log of the transaction price, whether the listing was an entire home, the log of the number of listings prior to April, 2014, and the log of the number of nights of the trip.

	Dependent variable:			
	Log(Nights in Exp.)	Log(Trips in Exp.)	Log(Nights by 2015)	Log(Bookings by 2015)
	(1)	(2)	(3)	(4)
First Treatment	−0.005	−0.003	−0.012*	−0.007
	(0.004)	(0.002)	(0.006)	(0.005)
Constant	0.279***	0.155***	2.193***	0.928***
	(0.003)	(0.001)	(0.004)	(0.004)
Observations	115,157	115,157	115,157	115,157
R^2	0.00001	0.00002	0.00003	0.00002
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 8: Long-term Guest Outcomes

A Robustness to Logging Inconsistencies

The simultaneous reveal review experiment launched on May 8, 2014 and included 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. Table AIII displays the baseline experimental results for this sample. The results between 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 1 or 2 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 4 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".[19] 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 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%. Figure A1 displays the most common phrases associated with negative reviews 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').

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

C Experimental Validity

This section documents that the experimental design in this paper is valid. Table AII displays the balance of observable characteristics in the experiments. There are no statistically significant differences in characteristics between the treatment and control guests or listings in the simultaneous reveal experiment.

D The Long-Run Evolution of Ratings

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

Figure A7 displays the review rates for guests and hosts over time, by treatment group. We see that following the end of the end of the experiment, when all groups were assigned the treatment, the review rates in the control groups quickly jump to match the review rates in the treatment group. This suggests that the longer exposure time for the treatment group did not have first-order consequences for reviewing behavior. It also suggests that the platform wide launch of the policy did not result in effects larger than those predicted by the experimental treatment effects on review rates.

Figures A8 and A9 display the long-run trends in the distribution of ratings by guests for a

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

E Stylized Model of Socially Induced Reciprocity in Reviews

In this section, we present a stylized model of the way in which reputation signals vary in the degree of bias they exhibit due to social aspects of the transaction. This simple model shows why expect to find the effect of socially induced reciprocity when the less salient review is negative and not when the less salient review is position.

The model is meant to illustrate the relationship between two signals, one of which is more affected by socially induced reciprocity (r_{2i}) and one of which is less affected (r_{1i}). For example, r_{1i} can be the recommendation and r_{2i} can be an indicator for a star rating greater than 3. Suppose, as in our regressions, that the reputation signals can take two values 0, 1. Furthermore, suppose that reviewers get utility from picking a signal that matches their experienced quality, q_i, and from sending more positive signals.

$$U(r_{1i}) = \mu_0 - (r_{1i} - q_i)^2 + \alpha_1 r_{1i} \tag{2}$$

$$U(r_{2i}) = \mu_0 - (r_{2i} - q_i)^2 + (\alpha_2 + \alpha_s s_i) r_{2i} \tag{3}$$

In the above equation, the terms α_i represent the preferences for sending more positive signals. We assume that $\alpha_2 > \alpha_1 > 0$ since the second signal is more salient to the host and more public.

We also assume that $\alpha_s > 0$. This is the term associated with the additional impact of socially induced reciprocity in transactions that have a more social component, s_i. We can use the above utilities to derive the reviewing functions.

$$r_{1i} = \begin{cases} 1 & \text{if } q_i > \frac{1-\alpha_1}{2} \\ 0 & \text{if } q_i \leq \frac{1-\alpha_1}{2} \end{cases}$$

$$r_{2i} = \begin{cases} 1 & \text{if } q_i > \frac{1-\alpha_2-\alpha_s s_i}{2} \\ 0 & \text{if } q_i \leq \frac{1-\alpha_2-\alpha_s s_i}{2} \end{cases}$$

Using these, and assuming a distribution for q_i, we can derive the conditional expectation, $E[r_{2i}|r_{1i} = j]$. Let $q_i \sim U[0,1]$ and assume that $\alpha_2 + \alpha_s \leq 1$.

Proposition 1:

$$E[r_{2i}|r_{1i} = 1] = 1 \tag{4}$$

$$E[r_{2i}|r_{1i} = 0] = \frac{\alpha_2 + \alpha_s - \alpha_1}{1 - \alpha_1} \tag{5}$$

The above proposition implies that the effect of socially induced reciprocity will be seen only when the less salient signal is equal to 0. Intuitively, when the reviewer rates the less salient signal positively, this means that the reviewer had a good experience and therefore, the more salient signal will be positive as well. In contrast, for a negative experience, the reviewer may submit a low r_{1i} and may shade r_{2i} upward due to socially induced reciprocity.

Our model abstracts from potential noise in the reviewing process. The addition of iid noise to the reviewing process will make it so that socially induced reciprocity will also affect, $E[r_{2i}|r_{1i} = 1]$, but to a relatively small extent if the amount of noise is small.

F Additional Tables

	Used iOS	Used Android	Used Desktop
	(1)	(2)	(3)
Treatment	0.010***	0.002**	−0.001
	(0.002)	(0.001)	(0.003)
Mean in Control	0.11	0.03	0.54
Has Recommendation in Control	0.11	0	0.51
Observations	119,789	119,789	119,789

Table AI: Effect of Experiment on Device Used

This table displays the effects of the experiment on whether a transaction had a review associated with each reviewing method. p<0.10, * p<0.05, ** p<0.01

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	

Table AII: Experimental Validity Check

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

	Host						Guest					
		All			*Conditional*			*All*			*Conditional*	
	Mean	Effect	SE	Mean	Effect	SE	Mean	Effect	SE	Mean	Effect	SE
Reviews	0.725	**0.067**	0.003				0.686	**0.009**	0.003			
Recommends	0.662	**0.061**	0.003	0.914	0.001	0.002	0.602	**0.017**	0.003	0.878	**0.012**	0.002
Overall Rating = 5							0.512	-0.003	0.003	0.746	**-0.014**	0.003
Lowest Subrating = 5	0.591	**0.047**	0.003	0.815	**-0.009**	0.003	0.329	-0.003	0.003	0.479	**-0.010**	0.004
Review Text Negative	0.019	**0.002**	0.001	0.027	0.001	0.001	0.084	**0.011**	0.002	0.123	**0.014**	0.003
Time to Review (Days)				3.784	**-0.389**	0.026				4.869	**-0.986**	0.042

Table AIII: Robustness to Alternative Sample

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. Values in bold are statistically significant at a 5% level.

	Dependent variable:		
	Non-recommendation and High Rating	Customer Service and High Rating	Low Rating and Positive Text
	(1)	(2)	(3)
Multi-listing Host	−0.0001	−0.002**	−0.002**
	(0.0003)	(0.001)	(0.001)
Trip, Guest, and Host Characteristics	Yes	Yes	Yes
Year-Month, Listing, and Guest FE	Yes	Yes	Yes
Mean of Y	0.0042	0.0325	0.0344
Obs	4937141	4937141	4937141
Note:			*p<0.1; **p<0.05; ***p<0.01

Table AIV: Multi-listing Host Status and Mismatch

The outcomes in the above regression are three measures of mismatch but more and less salient signals of transaction quality. The estimation is done on all trips with reviews between 2012 and 2014 for a 50% sample of guests. Additional controls for the number of prior reviews, nights of trip, guests in trip, number of prior bookings by the guest, and price per night are included. *p<0.10, ** p<0.05, *** p<0.01

G Additional Figures

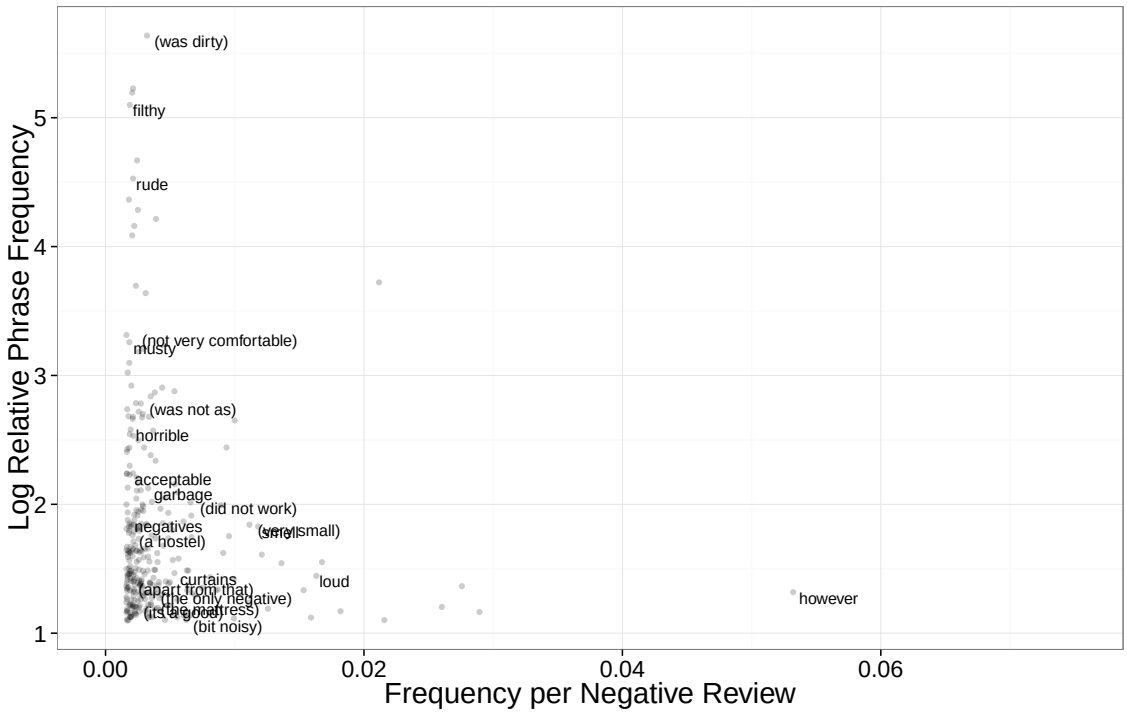


Figure A1: Distribution of negative phrases in guest reviews of listings.

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

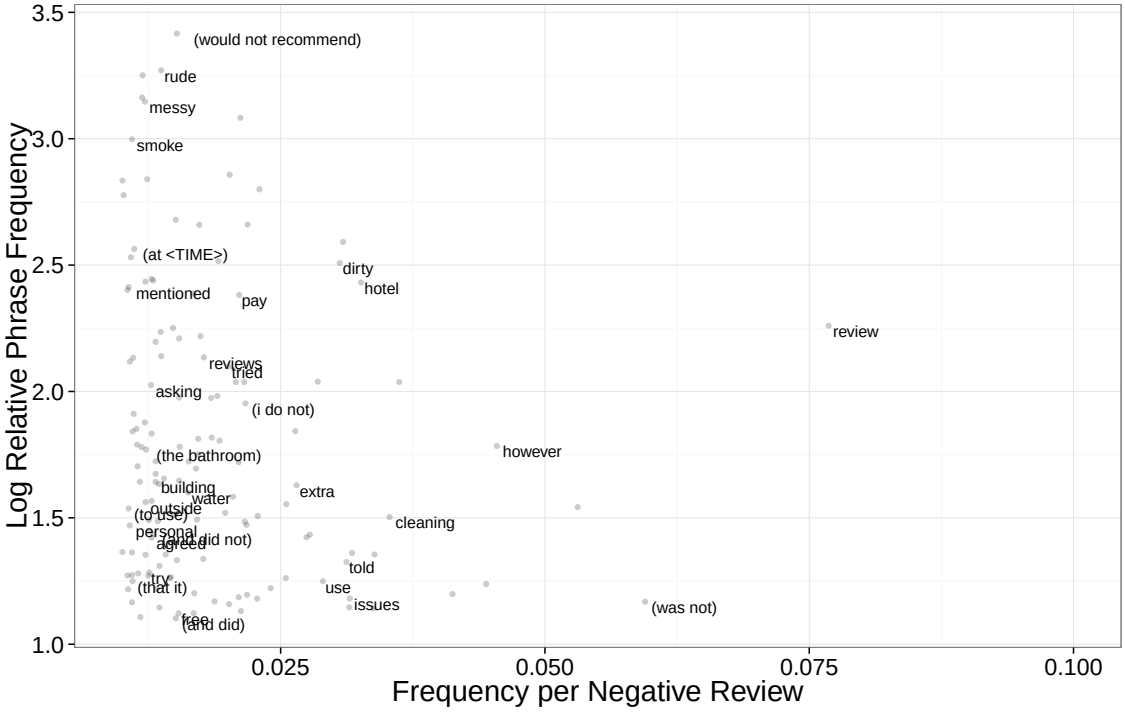


Figure A2: Distribution of negative phrases in host reviews of guests.

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

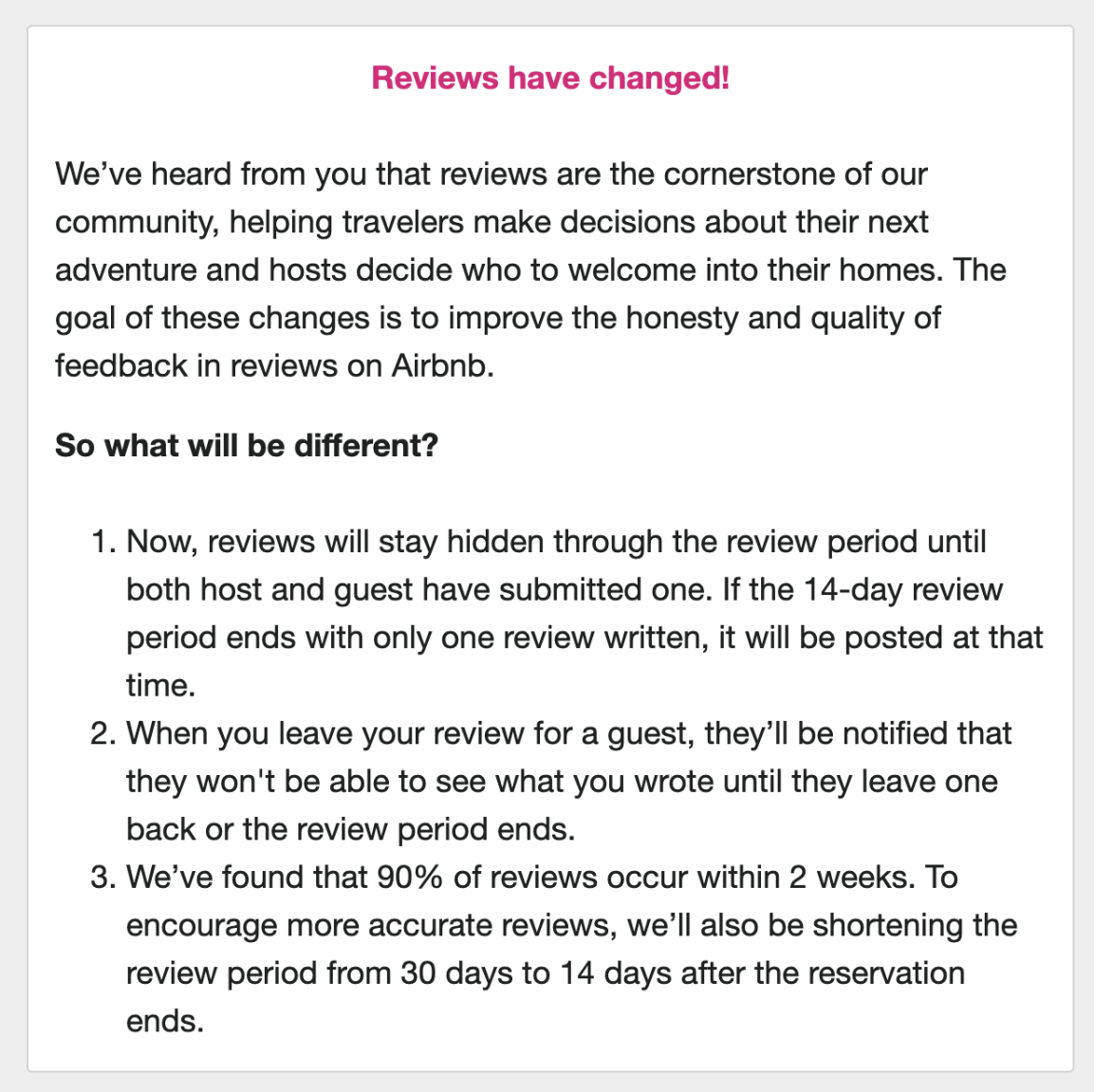


Figure A3: Interstitial in Some Host Emails

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

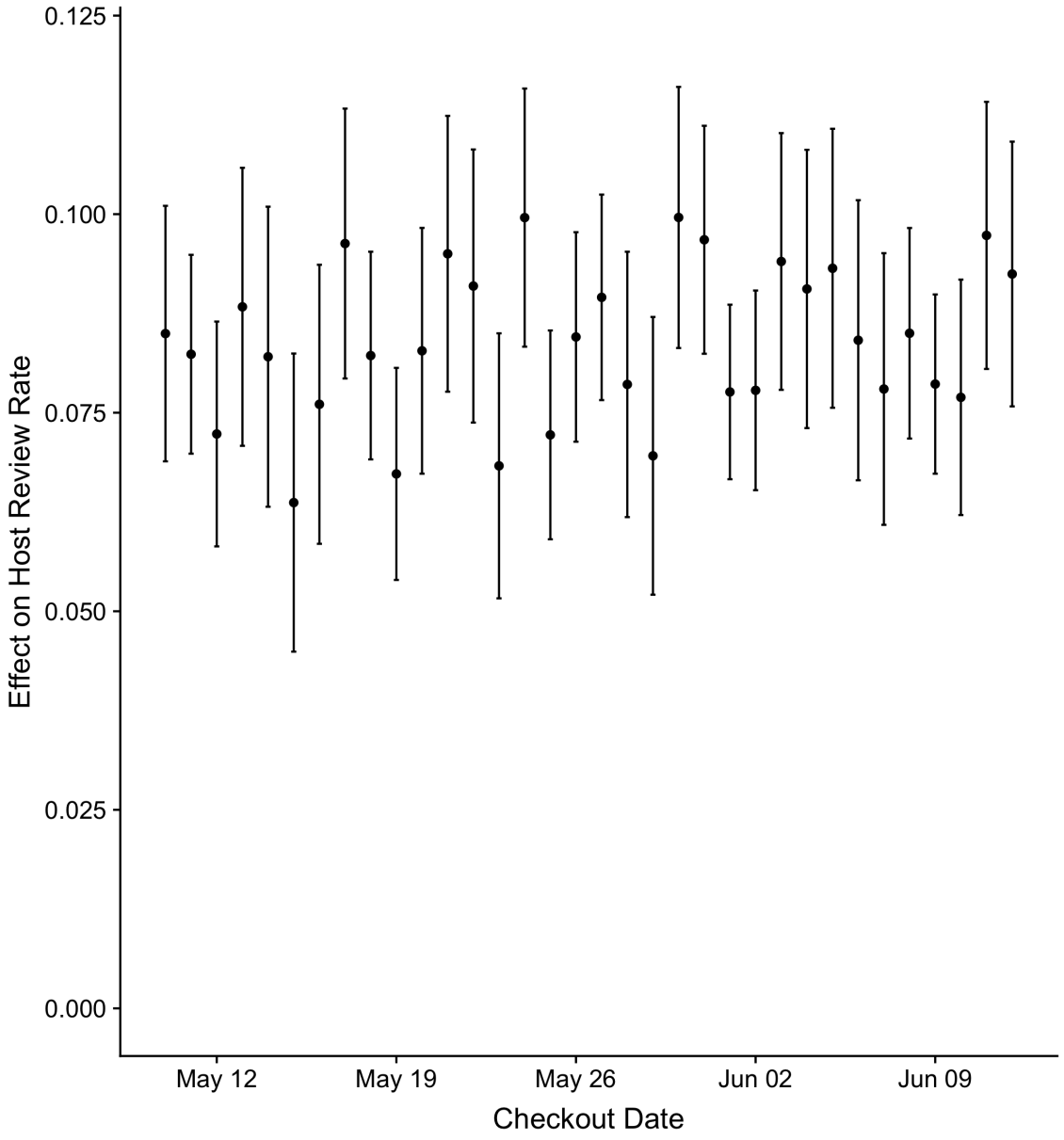


Figure A4: Effect of Treatment on Host Review Rates Over Time

This figure plots the daily treatment effect on host review rates and 95% confidence interval. To increase precision we use all transactions in the sample.

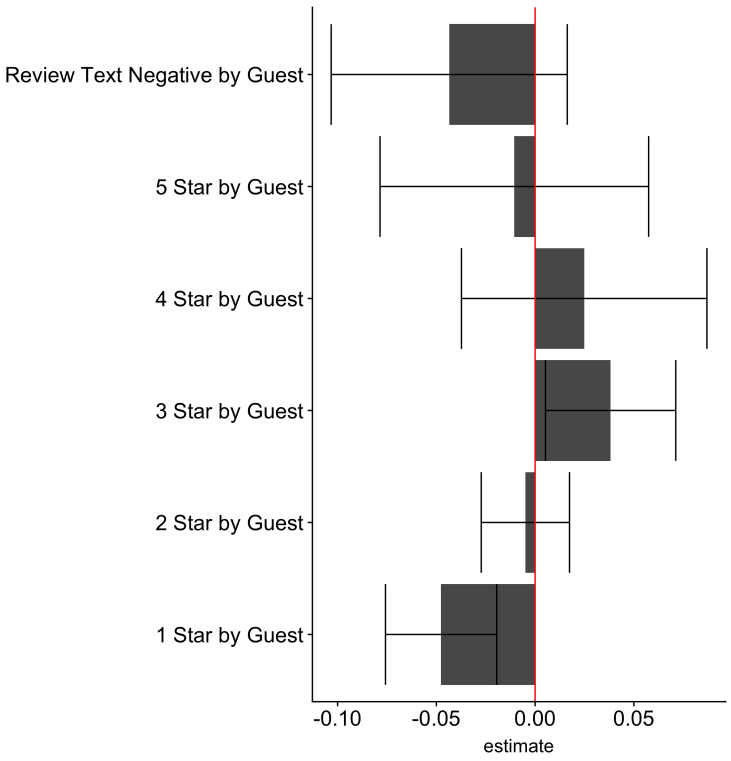


Figure A5: Differences in Guest Reviews Conditional on Negative First Host Reviews

This figure displays the difference between treatment and control in review content submitted by guests conditional on the host first submitting negative text. Note that this difference is not *causal* since it conditions on post-treatment outcomes.

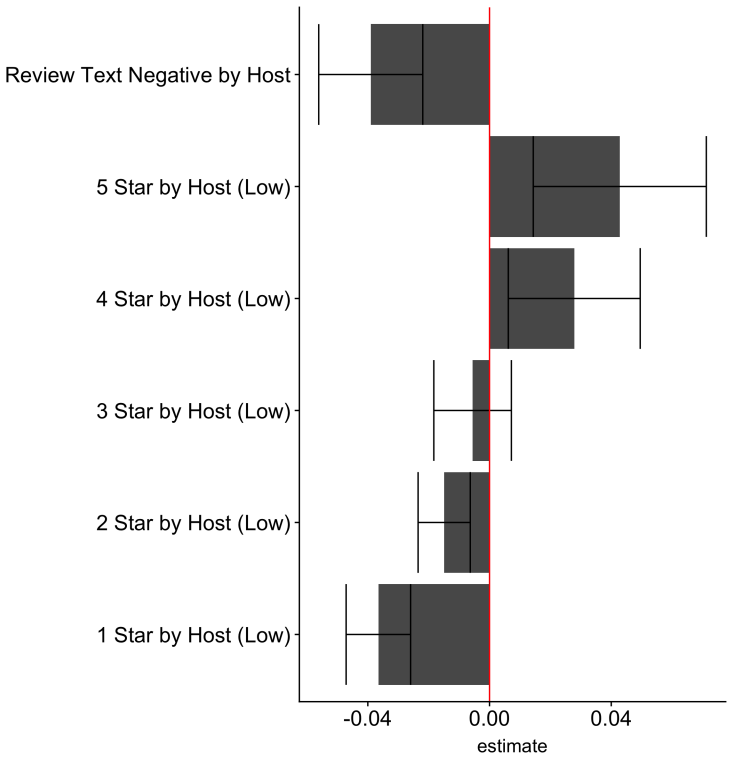


Figure A6: Differences in Host Reviews Conditional on Negative First Guest Reviews

This figure displays the difference between treatment and control in review content submitted by hosts conditional on the guest first submitting negative text. Note that this difference is not *causal* since it conditions on post-treatment outcomes.

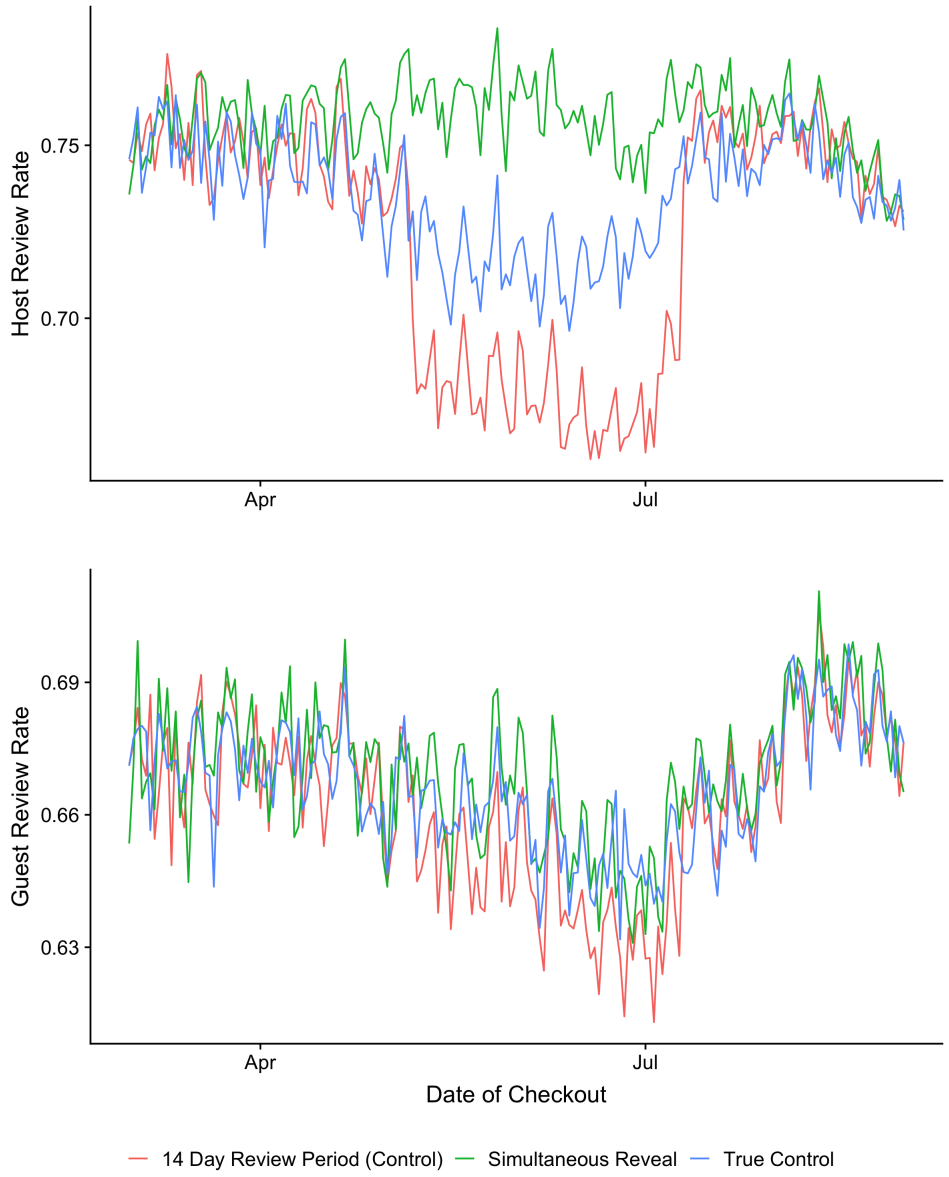


Figure A7: Review Rates Over Time

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).

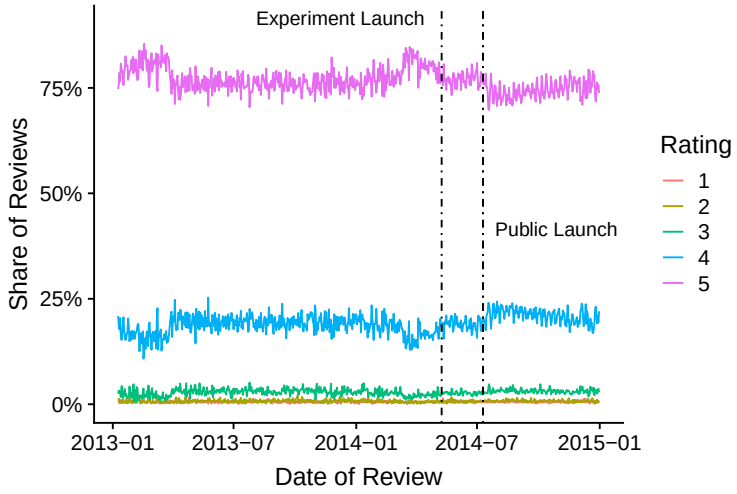


Figure A8: Ratings Over Time

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.

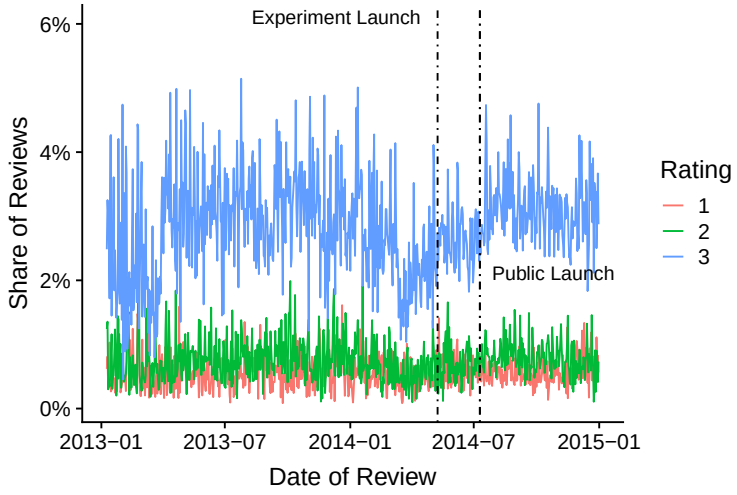


Figure A9: Ratings Over Time - Low Ratings

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.