The Determinants of Online Review Informativeness: Evidence from Field Experiments on Airbnb*

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

[1]MIT Sloan School of Management
[2]Airbnb, Inc.

April 20, 2018

Abstract

Reputation systems are used by most digital marketplaces, but their design varies greatly. We study how reputation system design affects the extent to which ratings and reviews aggregate information by using two experiments and complimentary observational analysis conducted on Airbnb. The first treatment that we study offered guests a $25 coupon in exchange for submitting a review. The second treatment implemented a simultaneous reveal review system, which eliminated strategic considerations from reviews. We show how both experiments made the reputation system more informative and use our findings to quantify the relative importance of mechanisms that cause inefficiency in reputation systems. We find that the largest source of inefficiency in reputation systems is sorting into reviews, whereby those who submit reviews typically have better experiences than those who do not. We also document retaliation and strategic reciprocity in the control group of our experiments but find that these mechanisms had small effects on the rating distribution. Lastly, we use observational analyses to document that social considerations also cause information loss in reputation systems. In summary, we find that reviews are typically informative but that negative experiences are underreported. We then discuss the implications of our findings for reputation system design.

*We are grateful to Chris Dellarocas, Liran Einav, Chiara Farronato, Shane Greenstein, John Horton, Caroline Hoxby, Ramesh Johari, Jon Levin, Mike Luca, Jeff Naecker, Fred Panier, Catherine Tucker, and seminar participants at Microsoft, MIT, eBay, HKU, ACM EC'15, NBER Summer Institute, and the CODE 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: afradkin@mit.edu
‡Primary Experiment Designer: elena.grewal@airbnb.com
§dholtz@mit.edu

1 Introduction

Reviews and ratings are used by nearly every digital marketplace and are widely considered to be critical for their success. These reputation systems reduce problems stemming from information asymmetry by soliciting information about transaction quality and displaying it to other market participants. However, the submission of accurate reviews is voluntary and costly, causing them to be underprovided in equilibrium (Avery, Resnick and Zeckhauser (1999); Miller, Resnick and Zeckhauser (2005)). This leads to missing information and a variety of biases.[1] Buyers who transact with sellers with biased reviews are less likely to transact on the platform in the future (Nosko and Tadelis (2015)). Consequently, the design of reputation systems that better aggregate information is an important challenge for digital platforms.

We use Airbnb, an online marketplace for short-term lodging, as the setting for the study of two reputation system experiments and complimentary observational analysis about reviewing behavior. We document three mechanisms that cause information loss on Airbnb, measure their relative importance, and quantify the overall informativeness of Airbnb reviews. Relative to previous studies of review systems, we are unique in studying multiple field experiments directly affecting the review system and on providing a methodology to quantify the importance of multiple forms of information loss in reviews. Furthermore, Airbnb is a part of a new sector often referred to as the "Sharing Economy." Reputation system design is particularly important in this sector because transactions are often social, involve heterogeneous services, and allow for two-sided feedback, creating complex reputation system design choices.

We begin the paper with a theoretical framework that motivates our subsequent analysis. In this framework, the choice of whether and how to review is influenced by a baseline utility of reviewing, an additional utility from reviewing positively, and the disutility from misreporting outcomes. The relative magnitudes of these factors vary across individuals, transaction types, and

[1] For a non-exhaustive list of documented biases in the literature, see these references: Dellarocas and Wood (2007); Cabral and Hortaçsu (2010); Saeedi, Shen and Sundaresan (2015); Bohnet and Frey (1999); Moe and Schweidel (2011); Nagle and Riedl (2014); Muchnik, Aral and Taylor (2013); Mayzlin, Dover and Chevalier (2014); Horton (2014).

reputation system designs, which we will experimentally manipulate. Buyers choose sellers based on observed reviews and ratings. The review system aids in market efficiency by identifying good and bad seller types. However, these review systems can lose information in two ways.[2] First, not everyone who transacts may review. Second, reviewers might not reveal their experiences in the public review. The realized information and efficiency loss from an imperfect system is a function of the agents' utility from reviewing, the design of the reputation system, and market conditions.

Next, we conduct an empirical analysis of Airbnb's review system and its informativeness. Throughout the paper, we use two notions of informativeness. The first one is whether a guest's review corresponds to the quality of their trip, which we proxy with private and anonymous recommendations about a guest or host submitted in the review flow and never shown to the transaction partner or other users. The reviewer should have less incentive to omit information about transaction quality in these recommendations. We find that these anonymous recommendations typically correspond to high ratings. The second piece of evidence is whether the review predicts outcomes plausibly related to the quality of an experience, such as whether a guest uses Airbnb again or whether there was a customer service complaint during the transaction. Indeed, higher ratings are associated with a higher likelihood of booking again and a lower likelihood of customer service complaints. Combining these two metrics, we show that recommendations have additional predictive power relative to star ratings and text when predicting future booking rates, meaning that some information is missing from the submitted reviews.

The above evidence for missing information in public reviews motivates our study of the causes of missing information in reviews, as well as interventions meant to recover this information. The first potential source of lost information in reputation systems that we consider is differential non-response. In particular, individuals with differing experiences may review at different rates (Dellarocas and Wood (2007); Nosko and Tadelis (2015)). We quantify differential non-response due to the quality experienced by the guest with an experiment conducted between April and July

[2]There is also considerable evidence about fake promotional reviews, which occur when firms post reviews either promoting themselves or disparaging competitors (see (Mayzlin, Dover and Chevalier, 2014) for a recent contribution). Promotional reviews are likely to be rare in our setting because a transaction is required before a review can be submitted.

of 2014. During the experiment, 50% of guests who stayed in a non-reviewed listing and had not left a review within 9 days of the trip's end received an e-mail offering a $25 dollar coupon in exchange for reviewing.

The treatment group experienced a 6.4 percentage point (pp) increase in review rates and the share of trips that were rated five stars increased by 2.4pp. Our analysis suggests that guests get more utility from leaving positive reviews and also dislike misrepresenting their experiences. These results corroborate findings from guest surveys conducted by Airbnb. Not only does the monetary incentive induce some of the guests to submit a review but the reviews those guests submit are accurate. Even in the treatment group of our experiment, not everyone reviews. We conduct a quantitative exercise to show that, were everyone to review, the five star rate would fall by at least 6pp.

Another potential source of bias in reviews is strategic reviewing behavior, where the reviewing decision of one party is affected by the predicted response of the other party in the transaction. There is evidence that strategic reviewing behavior may be especially important in the context of two-sided reputation systems (Cabral and Hortaçsu (2010); Saeedi, Shen and Sundaresan (2015)). For instance, observational studies comparing ratings across accommodation websites (e.g. Zervas, Proserpio and Byers (2015)) have led some to argue that Airbnb ratings are higher due to strategic considerations. Bolton, Greiner and Ockenfels (2012) propose a "simultaneous reveal" system in which reviews are hidden until both parties submit a review, and study the effects of this system in the lab. We conduct the first experimental test of such a simultaneous reveal mechanism in a live online marketplace.

Starting on May 8, 2014, Airbnb ran an experiment in which one third of hosts were assigned to a treatment that hid reviews until either both guest and host had submitted a review or a set period of time had elapsed. The treatment increased the review rate of guests by 1.8pp while decreasing the share of five star reviews by 1.6pp. 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 and increased the rate of negative sentiment in host review text conditional on a non-recommend from

71% to 74%. In summary, those in the treatment are more likely to review accurately and to submit negative reviews, but the overall effect is small.

We continue with a detailed analysis of the mechanisms behind the effects of the simultaneous reveal experiment. We show that there is lower correlation between guest and host reviews in the treatment. Next, for the set of trips where guests review second, we regress ratings by guests on ratings by hosts. We find that in the control, a host review with negative sentiment increases the likelihood of a guest's rating being below 5 stars. Most of this effect is eliminated in the treatment. Furthermore, hosts in the treatment group are 2.7pp more likely to review first, 8pp more likely to review first conditional on contacting customer support, and more likely to express negative sentiment in their written review. These results suggest that, in the control, hosts held back negative feedback either in fear of retaliation or to strategically induce reciprocity, and that the treatment mitigated this behavior.

A final potential source of bias in Airbnb reviews is the social nature of transactions, which has been shown to matter in lab experiments (Bohnet and Frey (1999); Sally (1995)). Stays on Airbnb frequently involve a social component, and internal Airbnb surveys of guests suggest that the social aspect of Airbnb affects reviewing behavior. Although we do not directly observe whether social interaction occurs, we observe proxy variables correlated with the degree of social interaction between guest and host. First, we observe whether the trip was to a private room within a home or to a private property; stays in a private room are more likely to result in social interaction because of shared space. Second, we observe whether a host is a multi-listing host (defined as a host with more than 3 listings) or a casual host; multi-listing hosts are less likely to interact with guests because they do not reside in the property. We use two separate identification strategies to study the causal effects of staying in private rooms and with multi-listing hosts. We show that, conditional on non-recommendations, there is a lower rate of high ratings for entire properties and multi-listing hosts. This evidence suggests that guests' willingness to write negative reviews is a function of the amount of social interaction they had with the host.

Each of the exercises above identifies a separate mechanism affecting reviewing behavior on

Airbnb. Next, we combine our estimates to measure the extent of information loss in the review system and to quantify the relative importance of these mechanisms. We begin by using the coupon experiment results to impute the rate at which ‘negative’ transactions occur and then use this rate to compute two measures of bias.

The first measure represents the difference between the true rate of negative experiences and the rate reported in reviews. There is a less than 2% difference between the imputed rate of negative experiences and the average rate of negative reviews. Sorting into reviewing is the most important source of this difference. The second measure of bias represents the share of negative experiences that were not reported in ratings. We find that 61% of negative experiences were not captured in ratings. Half of this effect is due to random non-response — there is some chance a guest will choose not to review, independent of the quality of their trip. The remaining 30% is explained by the three mechanisms described above. We use these results to calibrate our simple model of reviewing behavior and find welfare losses from an imperfect reputation system up to 3.1%, depending on market conditions.

Our results suggest that that the platform should focus on increasing review rates, especially for transactions with a higher likelihood of being negative. Furthermore, both of the treatments which we study have been implemented by the company. We discuss additional implications for reputation system design in the conclusion.

2 Setting

Airbnb describes itself as a trusted community marketplace for people to list, discover, and book unique accommodations around the world. Since 2008, Airbnb has seen over 300 million guest arrivals and has listed over four million places. Airbnb has created a market for a previously rare transaction: the rental of an apartment or part of an apartment in a city for a short term stay by a stranger.

In every Airbnb transaction that occurs, there are two parties - the “Host”, to whom the listing belongs, and the "Guest", who has booked the listing. After the guest checks out of the listing, there is a period of time (throughout this paper either 14 or 30 days) during which both the guest and host can review each other. Both the guest and host are prompted to review via e-mail the day after checkout. The host and guest also see reminders to review their transaction partner if they log onto the Airbnb website or open the Airbnb app. A reminder is automatically sent by email if a person has not reviewed within a given time period that depends on the overall review period or if the counter-party has left a review.

Airbnb's prompt for reviews of listings consists of two pages asking public, private, and anonymous questions (shown in Figure 1). On the first page, guests are asked to leave feedback consisting of publicly shown text, a one to five star rating,[3] and private comments to the host. The next page asks guests to rate the host 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 are displayed on each listing's page once there are at least three submitted reviews. Importantly, the second page also contains an anonymous question that asks whether the guest would recommend staying in the listing being reviewed.

The host is 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. The answers to these questions are not displayed anywhere on the website. Hosts also submit written reviews that will be publicly visible on the guest's profile page. Finally, the host can provide private text feedback about the quality of their hosting experience to the guest and to Airbnb.

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

3 Descriptive Statistics

In this section, we describe the characteristics of reviews on Airbnb. We use data for 59,981 trips between May 10, 2014 and June 12, 2014, which are in the control group of the simultaneous reveal experiment.[4] Turning first to review rates, 67% of trips result in a guest review and 72% result in a host review. Reviews are typically submitted within several days of the checkout, with hosts taking an average of 3.7 days to leave a review and guests taking an average of 4.3 days. Hosts review at higher rates and review first more often for two 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.

We first consider guest reviews of hosts. 97% of guests who submit a review for a listing, recommend that listing in an anonymous question prompt. This suggests that most guests report having a positive experience, even when there is no incentive to omit information. Figure 3a shows the distribution of star ratings for submitted reviews both conditional and unconditional on a recommendation. Guests submit a five star overall rating 74% of the time and a four star rating 20% of the time. 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. 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 lower than four star rating suggests that guests typically understand the review prompt.

Text comprises another important part of the review. We trained a logistic regression model to

[4]The experiments are randomized at a host level. Only the first trip for each host is included because the experimental treatment can affect the probability of having a subsequent trip. To the extent that better listings are more likely to receive subsequent bookings, these summary statistics understate the true rates of positive reviews in the website.

[4]There is no spike in the distribution for 1 star reviews, as seen on retail sites like Amazon.com. This is likely due to the fact that review rates are much lower for retail websites than for Airbnb.

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 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. 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 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. Last, reviewers may have accidentally clicked on the wrong rating.

Host reviews of guests are almost always positive. Over 99% of hosts responded that they would recommend a guest. These high ratings are present even though the prompt states: "This answer is also anonymous and not linked to you." Furthermore, only 14% of reviews by hosts have a category rating that is lower than five stars and less than 4% of reviews have negative text. We view this as evidence that the overwhelming majority of guests do not inconvenience their hosts beyond what is expected.

4 Theoretical Framework

In this section we describe a simple model of review mismatch and how reviewing behavior affects market efficiency. Suppose there is a marketplace that brings together buyers and sellers and that this marketplace operates for two periods. There are two types of sellers, a high type, h, and a low type, l. The low type sellers always generate a worse experience than the high type sellers. Each seller stays in the market for two periods and each period a mass of .5 sellers enter the market, with a probability, μ, of being a high type. Sellers choose a price, p ≥ 0 and their marginal cost is 0. Sellers do not know their type in the first period.

On the demand side, there is a mass of K identical buyers each period. Each buyer receives utility u_h if she transacts with a high type and u_l if she transacts with a low type. Furthermore, buyers have a reservation utility $\underline{u} > u_l$ and $\underline{u} \leq u_{\text{nr}}$, where u_{nr} is the minimum expected utility of sellers without reviews in periods 1 and 2. These assumptions ensure that buyers would not want to transact with low quality sellers but would want to transact with non-reviewed sellers. Lastly, after the transaction, the buyer can review the seller. Buyers can see the reviews of a seller but not the total amount of prior transactions.

After a transaction, buyers can choose whether and how to review sellers. Each buyer, i, has the following utility function for reviewing sellers:

$$\begin{aligned} \kappa_{ih} &= max(\alpha_i + \beta_i,\, \beta_i - \gamma) \\ \kappa_{il} &= max(\alpha_i + \beta_i - \gamma,\, \beta_i) \end{aligned} \tag{1}$$

where h and l refer to experiences with type h and l sellers respectively. β_i refers to the utility of submitting a review and is potentially influenced by the cost of time, financial incentives to review, and the fear of retaliation from a negative review. α_i refers to the additional utility of being positive in a review and is influenced by reciprocity and the general preference of individuals to be positive. γ is the disutility from misreporting. In the case of an interaction with a low quality seller, buyers have to make a choice between misrepresenting their experience, revealing their experience, or not reviewing at all.

Observation 1: Both types of sellers can have either no review or a positive review after a transaction. If $\beta_i < -\alpha_i$ then even if a guest transacts with an h seller, that guest will not leave a review. Furthermore, if α_i is high enough, then guests who transact with type l sellers may nonetheless leave a positive review.

Observation 2: The platform knows information that the buyers do not know about the likely quality of a seller. Since high type sellers are more likely to be reviewed in this setup, a non-review is predictive of the quality of a seller. The platform sees non-reviews while buyers do not and can use that information.

We now discuss the effects of changes to the parameters of the reviewing utilities, which correspond to our subsequent empirical exercises. Consider an increase in β_i, the baseline utility from reviewing, which is analogous to our subsequent coupon experiment. This change induces additional buyers to review but it does not change their decision to misreport conditional on reviewing. This can have opposing effects on review informativeness. First, increasing positive reviews of high type sellers and increasing negative reviews of low type sellers increases review informativeness. On the other hand, if most of the reviews on the margin are positive reviews of low type sellers, then this will actually decrease informativeness.

Decreasing α_i corresponds to a decrease in the relative utility of positive reviews. We think that this is one of the effects of both strategic and socially induced reciprocity. The change in review informativeness from decreasing α_i has three components. The first is that by inducing fewer positive reviews, high type sellers are identified later. The second is that low type sellers are less likely to be positively reviewed. The last is that expected type of non-reviewed sellers changes.

In the next section, we show that although reviews on Airbnb are informative, they are not perfectly informative. This means that there is room to increase the informativeness of Airbnb reviews through reputation system changes that affect the above parameters. In section 10, we revisit the above model to discuss the efficiency implications of increasing review informativeness in different settings.

5 Do Private Recommendations Contain Additional Information?

If reviews contain information about the quality of an experience, then they should predict verifiable measures of the quality of an experience and future usage of the platform by the reviewer. We demonstrate that reviews are indeed informative by showing that ratings predict future booking rates by guests and the presence of customer service tickets during the trip (in Appendix A) . Next, we show that the private and anonymous recommendations have additional predictive power over the public ratings and text. This is important to demonstrate because we use these recommendations throughout the rest of the paper as a measure of trip quality that should not be affected by strategic or social reciprocity.

Table 1 displays regressions where ratings are used to predict whether a guest books an Airbnb between August 2014 and May 2015. All specifications include controls for the prior experience of a guest because this is predictive of future bookings and use our entire sample. Column (1) shows that guests who review are 9.3pp more likely to book in the future. Column (2) adds controls for positive overall star rating, review text, and lowest category star rating. The overall star rating is the most informative, with an additional star being associated with a 2.3pp increase in re-booking rates. The lowest sub-rating is predictive even conditional on the overall rating, although the coefficient is smaller. Lastly, whether the review text is positive or not has no predictive value conditional on the ratings. Column (3) adds guest and trip characteristics such as number of nights, number of guests, and guest region. Even conditional on these characteristics, ratings continue to be predictive. Table AI displays similar results when the outcome is whether a customer service compliant occurred during the transaction.

So far, we've only used publicly visible review information in predicting future bookings. In column (4) of Table 1 we focus on cases where the public rating is high (greater than 3 stars) and look at the informativeness of the private and anonymous recommendations. Conditional on a star rating, a guest non-recommendation is associated with a 2.6pp decrease in re-booking rates.

Lastly, we investigate whether the predictive effect of ratings is driven by the types of listings that guests book. If all guests who stay at a given listing have similar re-booking rates regardless of rating, then it is likely that not every guest rates in accordance with their experiences. Alternatively, if guests who rate the same listing differently have different re-booking rates, then the ratings reflect heterogeneity of experiences during the stay or the preferences of a guest. Column (5) adds listing fixed effects to the above specifications. The coefficients on the rating related variables remain similar to specifications (2) and (3). Therefore, differences in ratings at least partially reflect differences in guests' experiences.

These results indicate that anonymous recommendations contain additional information not captured in star ratings, and that there is an opportunity for Airbnb reviews to be made more informative. This evidence for missing information in public reviews motivates our study of the causes of missing information in reviews, as well as the effectiveness of interventions meant to recover this information.

6 The Incentivized Review Experiment

6.1 Experimental Setup

In this section we document the results of the incentivized review experiment, which was intended to induce additional reviews for non-rated listings. The experiment was conducted between April and July of 2014 and consisted of all trips to non-reviewed listings for which the guest did not leave a review within 9 days. Among this sample, 50% of hosts were assigned to a treatment. Guests in the treatment group received an email offering a $25 Airbnb coupon in exchange for the review, while guests in the control group received a normal reminder email (shown in Figure 4).[5]

The treatment affected the probability of a review and the likelihood of a negative review. This resulted in more trips in the experimental sample to listings in the control group than listings in the

[5] The treatment email, like the control, did have the phrase "Hi ⟨name⟩," at the top, but it was cropped in the Figure 4.

treatment group. Therefore, we limit the analysis to the first trip to a listing in the experimental time-frame. Appendix C demonstrates that the randomization for this experiment is valid.

6.2 Overall results

Table 2 displays the review related summary statistics of the treatment and control groups in this experiment. First, note that the 26% review rate in the control group is smaller than the overall review rate (67%). The lower review rate is due to the fact that those guests who do not review within 9 days are less likely to leave a review than the average guest. The treatment increases the review rate in this sample by 70% and decreases the share of five star reviews by 8.0pp. The left panel of figure 2 displays the distribution of overall star ratings in the treatment versus the control. The treatment increases the number of ratings in each star rating category. It also shifts the distribution of overall ratings, increasing the relative share of 3 and 4 star ratings compared to the control. The non-public responses of guests are also lower in the treatment, with a 2pp decrease in the recommendation and likelihood to recommend Airbnb rates.

Table 3 displays the baseline treatment effects (Column 1) for this experiment using the sample of trips that were also in the treatment of the subsequent experiment (this sample is chosen for comparability of results).[6] The coupon treatment increases the review rate by 17pp and decreases the share of five star ratings by 13pp.

Column (2) displays the effect on the overall distribution of ratings taking into account not only stays in the experiment but stays which were reviewed within 9 days, which gives us the treatment effect on the overall ratings distribution (we describe the rescaling procedure in Appendix D). Rescaling the experimental treatment effect makes clear that the effect of the experiment on the overall distribution of reviews is smaller than that on the distribution of reviews in the experiment. One reason for this difference is that the sample of reviews in the experiment represent only a fraction of all reviews. Another reason for the difference between columns (1) and (2) is that guests

[6]Notably, the treatment effect of the coupon is larger when the simultaneous reveal treatment operates. The is likely due to the fact that without the possibility of retaliation guests face fewer costs from submitting negative reviews.

who review after 9 days tend to give lower ratings on average. Therefore, even if the experiment did not change the composition of reviews among those that did not review within 9 days, it would still have an effect on the distribution of ratings by inducing more of these guests to review. Put another way, not only does the monetary incentive provided by the coupon induce some guests to submit a review, but the reviews those guests submit are lower, reflecting their worse experiences.

6.3 Mechanisms

There are three potential reasons why the ratings in the treatment group are lower on average than the reviews in the control group. First, the coupon may have induced reviews for different types of transactions in terms of observables such as guest and listing characteristics. Second, the coupon may have induced those with worse experiences conditional on observables to submit reviews. Third, the presence of a monetary incentive may have changed the reviewing behavior of those that would have already left a review. We rule out explanation one and provide evidence that sorting on the quality of experience is one of the reasons for our effect.

First, we test whether the effect of the treatment on ratings persists after adding controls for a guest and listing characteristics. Columns (1), (2), and (4) of Table 4 show the treatment effects in specifications with progressively more controls. The treatment effects are similar across specifications. This rules out sorting on observables as an explanation for the effect.

Column (3) shows estimates of the treatment effect on ratings for a sample of experienced guests and adds controls for the historical leniency of a guest when submitting reviews. The guest leniency variable measures the extent to which the guest has previously submitted positive ratings. It is a binary variable equal to one if the guest specific fixed effect in a regression of ratings on guest fixed effects, along with listings / trip covariates, is greater than the median.[7] As expected, the coefficient on the guest leniency term is positive, with more lenient guests leaving higher ratings. The interaction between the treatment and whether the guest is lenient is not statistically significant and positive. This shows that guests do not substantially change their reviewing style as a result of the treatment, which would happen if the monetary treatment crowded out guest prosocial motivations. In column (5), we instead add an interaction between whether a host reviewed first and the treatment. The presence of a first host review induces reciprocity by the guest. If the coupon treatment reduces this reciprocity effect, then we would expect a negative coefficient on the interaction. Instead, we find a positive and statistically insignificant interaction effect. The results from these two specifications suggest that the coupon's effect comes from changing whether guests with worse experiences review rather than changing how guests rate conditional on reviewing.

An alternative way to test whether there is sorting on transaction quality is to look at the relationship between reviewing and guest rebooking rates. If the additional reviews occur due to sorting than those who review in the treatment group should have lower rebooking rates. However, this effect may not show up since reviewers in the treatment also received a coupon that should make them more likely to rebook.

Table 5 displays estimates of a linear probability model of whether a guest books between August 2014 and May 2015 as a function of the experimental treatment and whether the guest submits a review. First, there is no statistically detectable aggregate effect of the coupon on rebooking rates. Second, column (2) shows that those who submit reviews in the treatment group have lower rebooking rates than those who submit reviews in the control group. These effects are mitigated but persist even after including guest and listing characteristics in columns (3) and (4).

In summary, our results suggest that those who do not review in the control group have systematically worse experiences than those who do review. Furthermore, those not induced to review by the treament have even lower re-booking rates than those induced to review. To get a conservative estimate of the total bias due to sorting, we assume that non-reviewers had the same experiences on average as reviewers in the treatment group of the experiment. This allows us to predict the rating distribution if rating were compulsory. Column (5) of Table 3 displays the treatment effects under this scenario and shows that if everyone reviewed the five-star rate would be 6pp lower.

[7]The estimation sample for the fixed effects regressions is the year before the start of the experiment, so the estimated fixed effects are not affected by the experiment.

7 The Simultaneous Reveal Experiment

In this section we study the effects of a change in Airbnb's review system intended to remove strategic factors from reviews. 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 or reciprocated the first review. Furthermore, because of this possibility, first reviewers could strategically induce a reciprocal response by the second reviewer. To the extent that this behavior did not accurately reflect the quality of a reviewer's trip, it made the review system less informative.

The second experiment precludes these strategic reviews by changing the timing with which reviews are publicly revealed on Airbnb. Starting on May 8, 2014, Airbnb ran an experiment in which 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. The new user experience while reviewing is shown in Figure 5. 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 of 30 days to review.

For this analysis we limit the data we use to the first trip during the duration of the experiment to every listing that was in the experiment. We exclude subsequent trips because the treatment may affect re-booking rates, which would make the experiment unbalanced. Appendix C documents the validity of our experimental design. Table 6 shows the summary statistics for the treatment and control groups in the "simultaneous reveal" experiment. The treatment increased review rates for guests by 1.8pp. The rate of five star reviews by guests decreased by 1.6pp and the rate of guest reviews with negative text increased by 2pp. The treatment also induced a 6.4pp increase in the rate of guest suggestions to hosts. (see Table AVII). This increase was present even when conditioning on guest recommendations and star ratings. The relatively larger increase in private feedback rates suggests that without the fear of retaliation, guests felt they could speak more freely to the hosts about problems with the listing.[8]

[8]One worry about the external validity of these results is that not all guests and hosts may have noticed the infor-

Columns (3) and (4) of table 3 display the experimental treatment effects on guest reviews when controlling for trip and guest characteristics. Column (3) uses the entire experimental sample while column (4) shows estimates from a sample of previously non-reviewed listings. The effect sizes on ratings and sentiment are similar between the two samples, while the effect on reviews is smaller.

Turning to the host related statistics in Table 6, the rate of reviews increases by 7pp, demonstrating that hosts were aware of the experiment and were induced to review. Furthermore, the rate of positive recommendations by hosts increased by just 1pp, suggesting that the host recommendation is not substantively affected by the desire to retaliate. The text of the submitted reviews changes as well. The rate of negative sentiment conditional on a non-recommend (calculated using the methodology described in Appendix B) increases from 71% to 74%. This suggests that the experiment had the intended effect of inducing hosts to submit more informative public feedback. Table 7 displays a cross-tabulation of review ratings and text conditional on the treatment. The share of cases in which guests leave low ratings when hosts leave positive text increases by 3pp. Therefore, as expected, there is less correlation between guest and host reviews in the treatment than in the control.

Another way to look at the effects of the experiment is to see whether the treatment makes reviews more informative on average. To do this, we use review information to predict rebooking and customer service ticket rates (as in section 5), this time including an interaction between the rating and the treatment. The results are presented in Table AVI. Across specifications, the treatment does not significantly alter the relationship between ratings and outcome metrics. This is not surprising given that the experiment changed the distribution of ratings only slightly. Nonetheless, although ratings do not change in informativeness, the treatment results in more reviews and more reviews with lower ratings. Therefore, it increases the overall informational content of the review system.

---

mation about changes to the review system. We discuss this concern in Appendix E.

7.1 Evidence for Retaliation and Reciprocity

In this section, we use experimental variation to quantify the importance of strategic reciprocity in reviews on Airbnb. We first test for responses by the second reviewer to the first review. In the control group of the experiment, second reviewers see the first review and can respond accordingly. In the treatment group, second reviewers cannot respond to the content of the first review. Hence, the first review text should have no causal effect on the second review content, conditional on the host's recommendation. Our specification to test this is:

$$y_{gl} = \alpha_0 t_l + \alpha_1 FNR_{gl} + \alpha_2 FNS_{gl} + \alpha_3 t_l * FNR_{gl} + \alpha_4 t_l * FNS_{gl} + \beta' X_{gl} + \epsilon_{gl} \qquad (2)$$

where y_{gl} is a negative review outcome, t_l is an indicator for whether the listing is in the treatment group, FNR_{gl} is an indicator for whether the first reviewer did not recommend, FNS_{gl} is an indicator for whether the first review text contained negative sentiment, and X_{gl} are guest, trip and listing controls.

If guests reciprocate positive first reviews, then the guests in the treatment should leave less positive reviews after a positive review by a host. This response corresponds to α_0 being positive. Second, α_1 should be positive if there is positive correlation between guest and host experiences. Third, if there is retaliation against negative host reviews, α_2 should be positive because negative first review text induces negative second reviews. Moving to the interactions, α_2 should be negative because second reviewers in the treatment can no longer see the first review. Lastly, we expect that α_3, the interaction of the non-recommendation with the treatment to be close to 0. The reason is that second reviewers do not see the recommendation regardless of the experimental assignment.[9]

Table 8 displays estimates of Equation 2 for cases when the guest reviews second. Columns (1) - (3) show the estimates for guest non-recommendations, low ratings, and negative sentiment

[9]There are two complications to the above predictions. First, the experiment not only changes incentives but also changes the composition and ordering of host and guest reviews. If, for example, trips with bad outcomes were more likely to have the host review first in the treatment, then the predictions of the above paragraph may not hold exactly. Second, because we measure sentiment with error, the coefficients on the interaction of the treatment with non-recommendations may capture some effects of retaliation.

respectively. Turning first to the estimates of α_0, the effect is a precisely estimated 0 for non-recommendations and positive for the other metrics. This demonstrates that guests reciprocate positive reviews with positive public ratings but their anonymous ratings remain the same. Next, we consider the effect on a guest of a prior review by the host with negative sentiment, conditional on a non-recommendation. Across the three outcome variables, the coefficients on host negative sentiment range between .56 and .42. For instance, a host review with negative sentiment increases the likelihood of a guest's rating being below 5 stars by 42pp. This positive effect reflects a combination of retaliation and correlation in negative experiences between guests and hosts. The interaction of negative sentiment is of the opposite sign and ranges between -.33 and -.22. Therefore, at least some of the correlation between first and second negative reviews is driven by retaliation.

7.2 Evidence for Fear of Retaliation and Strategically Induced Reciprocity

We now investigate whether first reviewers strategically choose review content to induce positive reviews and to avoid retaliation. Strategic actors have an incentive to omit negative feedback from reviews and to wait until the other person has left a review before leaving a negative review. Because the simultaneous reveal treatment removes this incentive, we expect a higher share of first reviewers to have negative experiences and to leave negative feedback conditional on having a negative experience. We test for these effects using the following specification:

$$y_{gl} = \alpha_0 t_l + \alpha_1 NE_{gl} + \alpha_2 NE_{gl} * t_l + \beta' X_{gl} + \epsilon_{gl} \tag{3}$$

where y_{gl} is a negative review outcome, t_l is an indicator for whether the listing is in the treatment group and NE_{gl} is a measure of a negative experience (customer support ticket, non-recommendation, or negative sentiment). We expect α_0 and α_2 to be positive because first reviews should be more informative in the treatment.

Table AIII displays estimates of Equation 3 for first reviews by hosts. Column (1) shows that hosts are 2.7pp more likely to review first in the treatment. This demonstrates that hosts change their timing of reviews to a greater extent than guests. This likely occurs because hosts have more to lose from negative reviews and because hosts in the control group have an incentive to delay reviewing and implicitly threaten guests with a retaliatory review. Column (2) uses an indicator for whether the host contacted customer support as a measure of a negative experience. Hosts in the treatment were 8pp more likely to review first when they did contact customer support. Column (3) displays results when the outcome is negative text sentiment and the measure of a negative experience is a host non-recommendation. Hosts in the treatment are more likely to leave negative sentiment in a first review in the treatment, although this effect is not statistically significant. Lastly, the outcome in column (4) is a count of the number of negative words or phrases used in a review.[10] Conditional on leaving a negative textual review, hosts leave an additional .2 negative words in the review. These results demonstrate that hosts are aware of strategic considerations. In the control, hosts justifiably omit negative feedback from public reviews if they have a negative experience, either in fear of retaliation or to strategically induce reciprocity. The treatment mitigates this behavior.

8 Misreporting and Socially Induced Reciprocity

Reviewers leave conflicting private and public feedback even when there is no possibility of retaliation. In the simultaneous reveal treatment, guests who do not recommend a listing fail to leave text classified as negative 26% of the time and leave four or five star ratings 21% of the time. Similarly, hosts do not leave negative text in 26% of cases when they do not recommend the guest. In this section, we link some of this misreporting in public reviews to the type of interaction between the guest and host.

Stays on Airbnb frequently involve a social component. Guests typically communicate with hosts about the availability of the room and the details of the check-in. Guests and hosts also often

[10]Negative words or phrases are defined as those words or phrases that appear in at least 1% of non-recommend reviews and are at least 3 times more likely to appear in non-recommend than recommend reviews.

socialize while the stay is happening. This social interaction can occur when hosts and guests are sharing the same living room or kitchen. Other times, the host might offer to show the guest around town or the guest might ask for advice from the host. Lastly, the type of communication that occurs may differ between hosts who are professionals managing multiple listings and hosts who only rent out their own place. Internal Airbnb surveys of guests who did not leave a review suggest that the social aspect of Airbnb affects reviewing behavior.[11]

We do not directly observe whether social interaction occurs, but we do observe variables correlated with the degree of social interaction between guest and host. First, we observe whether the trip was to a private room within a home or to a private property; stays in a private room are more likely to result in social interaction because of shared space. Second, we observe whether a host is a multi-listing host (defined as a host with more than 3 listings) or a casual host. Multi-listing hosts are less likely to interact with guests because they manage many properties and typically do not reside in the properties they manage.

Figure 6 plots the distribution of guest ratings conditional on not recommending the host as a function of property type. Guests staying with casual hosts (those hosts that manage three or fewer listings) are over 5.5pp more likely to submit a five star overall rating than guests staying with multi-listing managers. That is, even though all guests in the sample would not recommend the listing they stayed at, those staying with multi-listing hosts were more likely to voice that opinion in a review rating.

However, this difference 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 and this could cause differences in the rates of misrepresentation. To control for these factors, we use two forms of variation in the data. First, 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. Second, stays with

[11] For example, one guest said: "I liked the host so felt bad telling him more of the issues" and another said "I often don't tell the host about bad experiences because I just don't want to hurt their feelings".

a multi-listing host may differ for a variety of reasons unrelated to socially induced reciprocity. Therefore, we use variation within listing to study the effects of multi-listing hosts. Specifically, because hosts sometimes start as casual hosts but then expand their operations over time, we observe stays at the same listing as the host transitions from a casual to a multi-listing host. This allows us to identify the effect of multi-listing host status on guest reviews.

Consider the following regression specification:

$$y_{glt} = \alpha_0 PR_l + \alpha_1 MLH_{lt} + \alpha_2 R_{glt} + \alpha_3 MLH_{lt} * R_{glt} + \alpha_4 PR_{lt} * R_{glt} + \beta' X_{gl} + \gamma_g + \epsilon_{gl} \quad (4)$$

where y_{glt} is a negative review by guest g for listing l at time t, PR_l is an indicator for whether the listing is a private room, MLH_{lt} is an indicator for whether the host is a multi-listing host, R_{gl} is an indicator for whether the guest recommends the listing, X_{gl} are guest and trip characteristics, and γ_g is a guest fixed effect. If socially induced reciprocity occurs, then α_3 should be negative because guests to private rooms should leave less negative feedback and α_4 to be positive because multi-listing hosts induce less reciprocity in guests.

Table 9 displays the results of regressions predicting whether a review rating had more than 3 stars. Column (1) contains a specification with controls for guest, trip, and listing characteristics, while column (2) adds guest fixed effects. In both specifications, entire properties are 1 percentage point less likely to receive high rating, but the effect goes away if a guest recommends a listing. Similarly, multi-listing hosts are 4.5pp less likely to receive high rating when they are not recommended by the guest. Column (3) adds listing fixed effects, using variation in host status over time to identify the effect of a multi-listing host. In this case, reviews of multi-listing hosts are 3.2pp less likely to receive high rating if the guest does not recommend.

Table AVIII contains the specifications with address fixed effects. Column (1) shows a regression in which the entire property indicator is not interacted with the recommendation. There is no difference on average between reviews of entire properties and private rooms at the same location. However, when interactions are added in columns (2) and (3), there is 4.6pp decrease in the probability of high ratings for entire properties relative to private rooms conditional on a non-recommendation. This evidence confirms that guest's willingness to write negative public reviews is a function of the degree of social interaction they had with the host. These estimates of socially induced reciprocity are likely to be underestimates because even stays at entire properties with multi-listing hosts still sometimes have a social component.

9 Measuring the Size of Bias

Our analysis has shown that submitted reviews on Airbnb are not fully representative of transaction quality due to sorting, strategic reciprocity, and socially induced reciprocity. In this section, we describe a methodology for using experimental estimates and observational data to measure the extent of information loss in a review system and to quantify the relative importance of the mechanisms documented in this paper.

9.1 Empirical Analogues of Bias Measures

Suppose that each trip results in a positive experience with probability, g, and a negative experience (denoted n) with probability, 1 - g. An unbiased review system would have a share, g, of positive ratings. Furthermore, suppose that there are only two types of reviews, positive (s_g) and negative (s_n). Then the share of submitted ratings that are positive is:

$$\bar{s} = \frac{gr_p + (1-g)r_{lp}}{Pr(r)} \tag{5}$$

where r is an indicator for whether a review was submitted, $r_p = Pr(s_g|p)$, is the probability of a positive review after a positive stay, $r_{lp} = Pr(s_p|n)$, is the probability of a positive review after a negative experience, $r_{ll} = Pr(s_n|n)$ is the probability of a negative review after a negative experience, and $Pr(r) = gr_p + (1-g)(r_{lp} + r_{ll})$ is the overall review rate. The difference between the average actual experience and the average submitted review is:

$$B_{avg} = (1-g)\frac{r_{lp}}{Pr(r)} - g(1 - \frac{r_p}{Pr(r)}) \tag{6}$$

The first term is the share of reviewers who have negative experiences and report positively and the second term is the share of reviewers who have positive experiences but do not report positively. Note, these two forms of bias push the average in opposite directions so that looking at average ratings understates the amount of misreporting.

Our second measure of bias is the share of negative experiences not-reported by reviewers:

$$B_{neg} = 1 - \hat{r}_{ll} = 1 - \frac{N_{n|n}}{N_{all}(1-g)} \tag{7}$$

where $N_{n|n}$ is the number of negative reports given the reviewer has a negative experience and N_{all} is the total number of trips.

In order to operationalize these metrics, we assume that guests reveal the quality of a transaction in the anonymous recommendation. To calibrate the empirical analogue to g, we need to make assumptions about the degree of selection into reviewing. First, note that the recommendation rate for guests in the incentivized review experiment was lower than in the control, $Pr(r|g) \neq Pr(r|b)$. Therefore, we cannot simply use the rates of recommendations in the data to back out g. Instead, we calibrate g by using the recommendation rates from the incentivized review experiment, which eliminates some of the effect of selection into reviewing. However, because this experiment was only conducted for listings with 0 reviews, we must extrapolate to the sample of all reviews.

To calibrate $\hat{g}$ we need to make two assumptions about reviews. First, we set the rate of positive experiences for those that do not review when offered the coupon equal to the rate of positive experiences for guests eligible for the coupon experiment who reviewed in the treatment group of the experiment. This assumption is conservative, given that those not induced to review by the Airbnb coupon are likely to have even worse experiences on average than those that did review. Second, we must make an assumption about trips to reviewed listings, which were not eligible for the incentivized review experiment. We assume that the relative rate of bias due to sorting must be the same across listings with different amounts of reviews. In the absence of experimental variation, we cannot confirm or reject this proposition. We then reweigh the baseline rate of recommendation for listings with 0 reviews by the relative rates of recommendations in the overall sample (see Appendix D for details). Lastly, we need to calibrate the review probabilities and misreporting rates conditional on leaving a review. We describe how to do so in the next section.

9.2 The Size of Bias

We measure bias for guest reviews of listings in five scenarios, each with progressively less bias. Scenario 1 represents the baseline scenario in the control group in the simultaneous reveal experiment. In this case all three mechanisms (sorting, strategic, and social) operate. Scenario 2 corresponds to the treatment group of the simultaneous reveal experiment (note, there are effects on both ratings and review rates). In both scenarios, we calculate measures of bias by making transformations of the moments in the data. $\widehat{Pr(s_g|n, r)}$ is equal to the empirical rate of positive reviews (either a high rating or positive text) without a recommendation. $1 - \hat{g} = 3.68\%$ is our estimate of the true rate of negative experiences and $\widehat{Pr(r|n)} = \frac{\widehat{Pr(n|r)} * \widehat{P(r)}}{(1-\hat{g})}$. Combing these, $\hat{r}_{lp} = \widehat{Pr(s_g|n, r)} * \widehat{Pr(r|n)}$. Scenario 3 represents the bias if there was no socially induced reciprocity in the reviewing process. To calculate the review rates in this scenario, we set $\widehat{Pr(s_g|n, r)}$ equal to the adjusted rate of positive reviews for stays with multi-listing hosts in entire properties. Scenario 4 turns off the sorting mechanism. This corresponds to the scenario where the rate of non-recommendation reviews is equal to the share of guests with negative experiences but the overall review rate was equal to the review rate in the simultaneous reveal treatment. Lastly, scenario 5 computes the two measures of bias if everyone submits reviews.

Table 10 displays both measures of bias in each of the five scenarios.[12] We first turn to the case when all biases are present (row 1). In this scenario, positive reviews occur 1.32% more frequently than positive experiences. Furthermore, 61% of non-recommended experiences are not reported in ratings. Rows 2 and 3 display the effects of removing strategic and social reciprocation. Removing these mechanisms reduces the average bias by .18pp and the share of negative reviews missing by 4.6pp. Therefore, both strategic and social reciprocity account for a relatively small portion of the bias in the system.

In row 4, we remove sorting bias. There is a fall of 1.1pp in average bias and 25pp in the share of negative experiences missing. This shows that sorting is a more important source of bias than strategic and socially induced reciprocity using both measures. In row 5 we report what our measures of bias would be if every guest submitted a review conditional on removing the aforementioned biases. In this case, B_{avg} does not change because the rate of misreporting does not change. However, B_{neg} falls by an additional 31pp due to the fact that even without sorting into reviewing, some non-reviewers would have negative experiences which would not be reported. Lastly, there is a residual 1.1% of negative experiences that would still go unreported. This is due to misreporting and can correspond to two scenarios: measurement error or residual socially induced reciprocity that occurs even when guests stay at the properties of multi-listing hosts.

The average level of bias on this site is small if we only care about a binary signal of quality. In total, there is a less than 2% difference between the imputed rate of negative experiences and the average rate of negative reviews. Of the negative experiences that did occur, 61% were not captured in ratings. Bias due to both strategic and social reciprocity exist, but comprise a relatively small share of the total. Sorting into reviewing represents the biggest source of bias on the platform. Furthermore, even without sorting or reciprocity, approximately 30% of guests do not submit a review due to random non-response – there is some chance a guest will choose not to review, independent of the quality of their trip. This constitutes a final source of missing information in the reputation system.

[12] See Table AV for a measure of bias using text sentiment conditional on a non-recommendation. The results are a qualitatively similar, although both measures of bias are higher due to the fact that there is more mismatch between review text and recommendations than review ratings and recommendations.

10 The Efficiency Implications of Our Experiments

Now that we have quantified the amount of information loss resulting from sorting, strategic reviewing behavior, and socially induced reciprocity, we can revisit our theoretical framework and analyze the efficiency implications of reputation system design, both in the general sense and in the particular setting of Airbnb. To simplify the analysis, we refer to the review probabilities (r_{lp}, r_{ll}, r_p) rather than review utilities. Furthermore, we assume that the review rates satisfy the following inequality, $(1-\mu)r_{lp}u_l + \mu r_p u_h > u_{\text{nr}}$, so that positive reviews increase the expected utility from a trip, a natural assumption in our setting. Lastly, let $\bar{u} = \mu u_h + (1-\mu)u_l$, the period 0 expected utility from a booking.

Consider first the case where K > 1, meaning that there are many more buyers than sellers. This, corresponds to times of the year when there is a large travel demand and markets where there are relatively few hosts. In this scenario, all sellers without a negative review transact because buyers' expected utility from the transaction is higher than their reservation utility. The surplus from having the reputation system in period 2 (excluding the disutility from reviewing) is:

$$S_{\text{Status Quo}} = \bar{u} + .5r_{ll}(1-\mu)(\underline{\text{u}} - u_l) \tag{8}$$

Now suppose that everyone reviewed and fully revealed their experience. This corresponds to $r_p = 1$ and $r_{ll} = 1$. The difference in surplus between this scenario and the status quo is $.5(1-\mu)(\underline{\text{u}} - u_l)(1 - r_{ll})$. We can use our previous analysis of bias in the reputation system to calibrate the above expression. First, note that $B_{neg} = 1 - r_{ll} = .61$. Second, while we do not directly observe μ, we know that it's less than $\hat{g} = .964$ and that most guests staying even with non-reviewed listings have a good experience. Consequently, μ must be high. Lastly, we do not measure $\underline{\text{u}} - u_l$ but this represents how much worse a bad experience is than the outside option, which would typically be a hotel. If we assume that $\mu = .95$ and $\underline{\text{u}} - u_l = -u_l = -\$157.50$ (the median price per night in our sample), then an imperfect reputation system reduces efficiency per night by \$4.87 and removing sorting and reciprocity reduces that to \$2.50.[13] This represents 3.1%

of the purchase price, which can sum to hundreds of millions of dollars when aggregated across all transactions.

In contrast, the gains from a better reputation system are different when the ratio of buyers to sellers is low, such as during low-demand periods in low-demand cities. Let K $< .5\mu r_p$, meaning that there are fewer buyers than reviewed high quality sellers. The surplus gain from having a reputation system in this case is:

$$\frac{S_{\text{Status Quo}}}{K} = \frac{\mu u_h r_p + (1-\mu) u_l r_{lp}}{\mu r_p + (1-\mu) r_{lp}} - \bar{u}, \; \frac{S_{\text{Perfect}}}{K} = u_h - \bar{u} \tag{9}$$

Consequently, the gains from a perfect reputation system are:

$$\frac{S_{\text{Perfect}} - S_{\text{Status Quo}}}{K} = (u_h - u_l)\frac{(1-\mu) r_{lp}}{\mu r_p + (1-\mu) r_{lp}} \tag{10}$$

In the above expression, what matters for efficiency is the number of positive reviews of low type sellers relative to the number of positive reviews of high type sellers. This occurs because all bookings are of reviewed sellers. Furthermore, the utility difference that matters is $u_h - u_l$ rather than $\underline{u} - u_l$. We can impute $\hat{r}_{lp} = .041$ and $\hat{r}_p = .70$. Lastly, the consumer surplus estimates from Farronato and Fradkin (2018) suggest that u_h is approximately 16% higher than the purchase price for those who transact on Airbnb. Combining our previous assumptions we can calculate that the loss per night in this setting equals \$1.10. One way to interpret this, is that in times where Airbnb has many reviewed sellers relative to buyers, the losses from an imperfect reputation system are relatively small. For intermediate cases of K, all three review rates, (r_p, r_{ll}, r_{lp}), matter for welfare.

[13] These calculations abstract away from dynamic considerations. For example, consumers may leave the platform for good and this may harm the platform even more. Furthermore, reviews serve to reduce moral hazard among sellers.

[13] See Table AV for similar results regarding review text.

11 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. This bias can reduce market efficiency in a variety of ways. In this paper, we use experiments and proprietary data from Airbnb to show that, at least in this setting, public reviews are informative and typically correspond with private and anonymous ratings.

Nonetheless, reviews are not fully informative and market design affects what information is revealed. We use experiments to study the effects of market design changes and to document three mechanisms which can distort online reputation: sorting, strategic reciprocity, and socially induced reciprocity. Our first experiment offers a coupon for guests to submit reviews of non-reviewed listings. We show that this experiment decreases the ratings of submitted reviews and document that this effect is caused primarily by sorting. Next, we study the simultaneous reveal experiment, which eliminates strategic reciprocity in reviews. We show that this experiment reduced five star ratings of hosts by guests by 1.5pp and increases review rates by 1.8pp. Lastly, we document that some of the remaining mismatch is caused by the social nature of the transaction. Altogether, we estimate that these sources of bias cause a less than 2pp difference between the rate of negative experiences and the rate of negative reviews. However, of those experiences that are negative, we estimate that 61% are not reported in review ratings.

Our results suggest that, for Airbnb's review system, the most important challenge to tackle is sorting into reviewing. Indeed, Airbnb has continued to experiment with changes intended to increase review rates. We find that the platform should be most concerned about an imperfect reputation system in markets with a high ratio of buyers to sellers. Our calibration of a simple model of reviews suggests that at these times, an imperfect reputation system can cost 3.1% of the transaction price in welfare and that removing the biases we identify would reduce this loss by over 50%. If one thinks of this efficiency loss as analogous to a cost margin, then improving reviews could give Airbnb a substantial competitive advantage. In contrast, when there are few buyers to sellers, the losses from an imperfect reputation are relatively small. A similar methodology to ours can be used to diagnose and improve reputation systems in other marketplaces, where the relative importance of these mechanisms has yet to be studied.

There are several interventions that might reduce sorting bias in addition to the incentivized review policy which we study. First, marketplaces can change the way in which reviews are prompted and displayed in order to increase review rates. For example, the simultaneous reveal experiment described in this paper increased review rates and consequently reduced the rate of sorting into reviewing. Other potential interventions include making reviews mandatory (as has previously been the case on Uber) or making the review easier to submit. Second, online marketplaces can display more informative reputation metrics in addition to simple averages and distributions of submitted reviews. For example, the effective positive percentage could be shown on a listing page in addition to the standard ratings. Alternatively, listing pages could be augmented with data on other signals of customer experience, such as customer support calls. Also, as in Nosko and Tadelis (2015), the platform can use its private information regarding the likely quality of a listing to design a search ranking algorithm. Each of these policies is promising but may have unintended consequences by affecting the likelihood that a submitted review is informative. We anticipate that our results will stimulate further research on cost-effective interventions that increase review rates without compromising review informativeness.

Lastly, our theoretical model suggests that a key variable determining the benefits from a reputation system is the share of high type sellers entering the platform. We've shown that this rate is high in Airbnb's system. This raises the question of why more low quality sellers do not enter. There are at least three possible mechanisms that may account for the proportion of high quality sellers on the platform. First, many bad actors or listings may be caught by Airbnb's trust and safety efforts. These efforts include verifying the identities of guests and hosts, tracking and preemptively eliminating scams, encouraging detailed profiles, and subsidizing high resolution photos. Second, the search ranking algorithm might explicitly reduce the rankings of low-quality sellers. Third, the law of large numbers may ensure that low quality listings are eventually negatively reviewed and consequently never booked again. We leave the study of these mechanisms for future work.

References

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. "Social Distance and Other-Regarding Behavior in Dictator Games: Comment." *American Economic Review*, 89(1): 335–339.

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.

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."

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

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.

Muchnik, Lev, Sinan Aral, and Sean J Taylor. 2013. "Social influence bias: A randomized experiment." *Science*, 341(6146): 647–651.

Nagle, Frank, and Christoph Riedl. 2014. "Online Word of Mouth and Product Quality Disagreement." Academy of Management Proceedings SSRN Scholarly Paper ID 2259055, Rochester, NY.

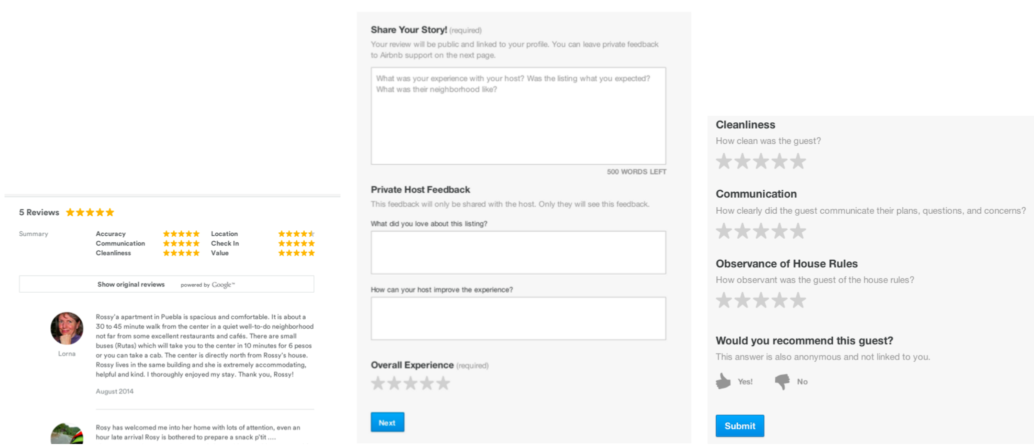
Nosko, Chris, and Steven Tadelis. 2015. "The Limits of Reputation in Platform Markets: An Empirical Analysis and Field Experiment." *NBER Working Paper*, , (20830).

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

Sally, David. 1995. "Conversation and Cooperation in Social Dilemmas A Meta-Analysis of Experiments from 1958 to 1992." *Rationality and Society*, 7(1): 58–92.

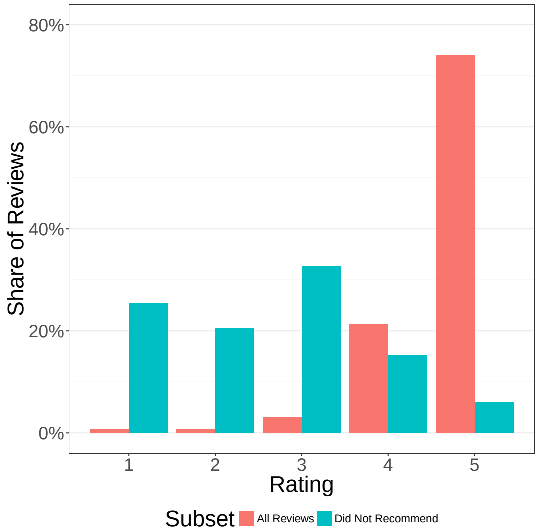
Zervas, Georgios, Davide Proserpio, and John Byers. 2015. "A First Look at Online Reputation on Airbnb, Where Every Stay Is Above Average." Social Science Research Network SSRN Scholarly Paper ID 2554500, Rochester, NY.

12 Figures

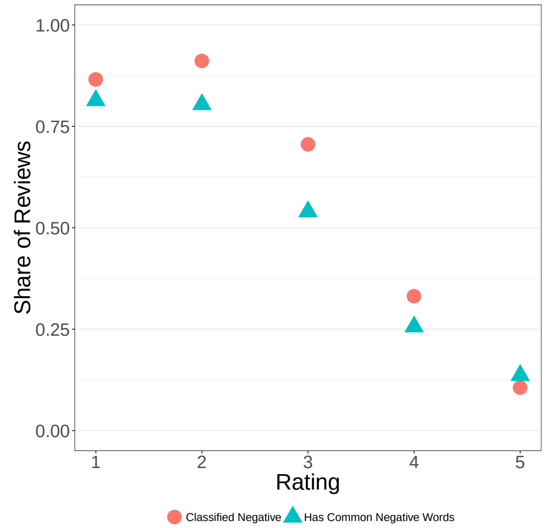


(a) Reviews on Listing Page (b) Review of Listing (Page 1) (c) Review of Guest (Page 2)

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.

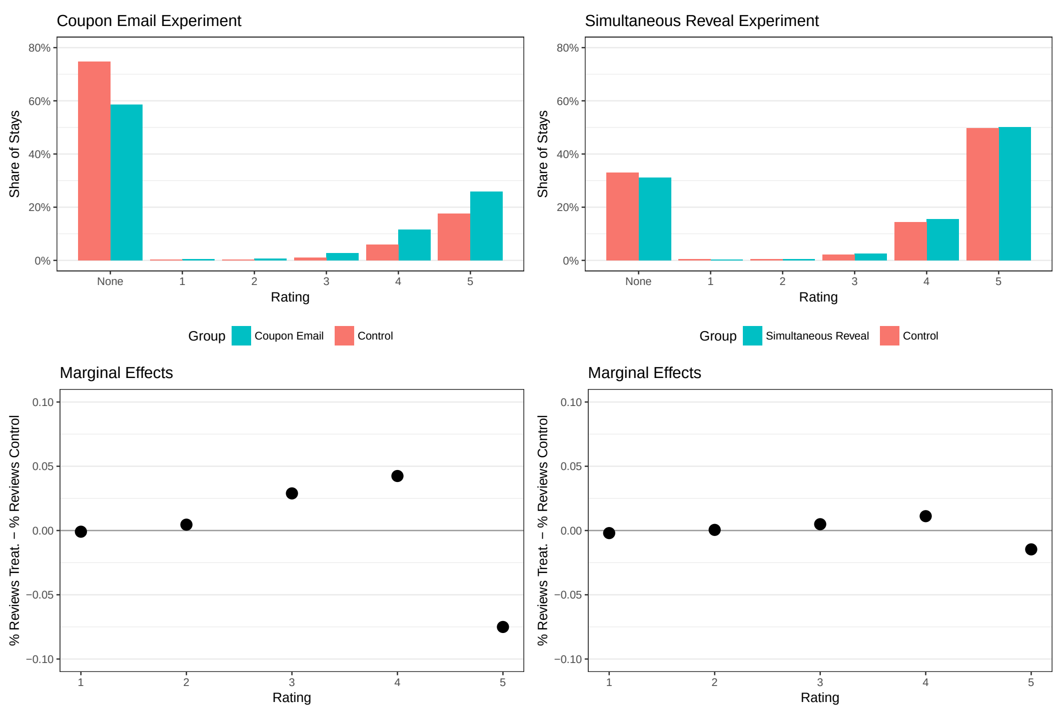
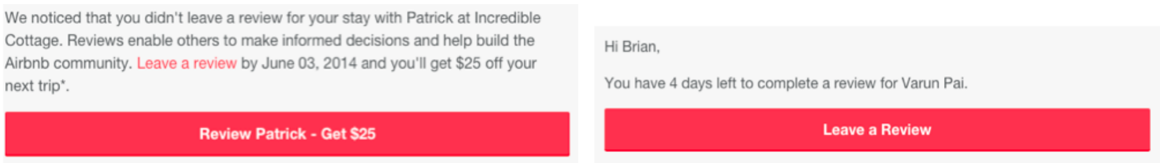


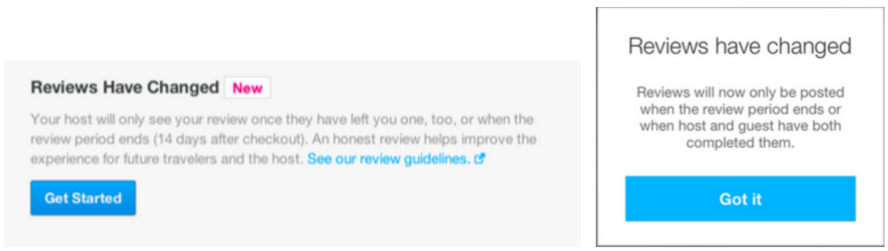
Figure 3: Distribution of Ratings - Experiments

The above figure displays the distribution of ratings in the control and treatment groups for the simultaneous reveal experiment and for the incentivized review experiment. Row 1 displays the distribution of reviews while row 2 shows the marginal effects on ratings.



(a) Treatment Email (b) Control Email

Figure 4: Incentivized Review Experiment Emails



(a) Desktop (b) Mobile

Figure 5: Simultaneous Reveal Notification

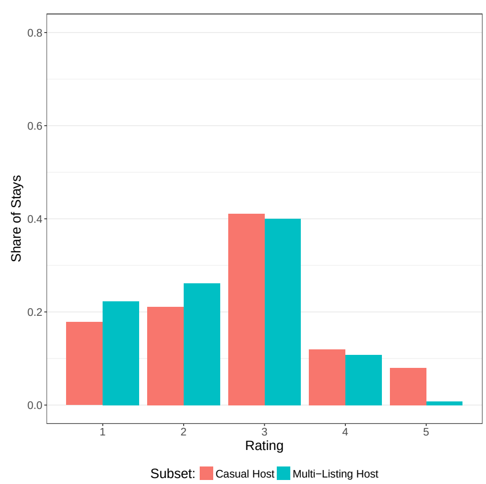


Figure 6: Ratings When Guest Does Not Recommend

The above figure displays the distribution of submitted ratings in the treatment group of the simultaneous reveal experiment. This is done to remove strategic factors from also affecting the distribution. Only reviews for which the guest anonymously does not recommend the host are included. Only the first stay for each listing in the experiment is included.

13 Tables

	Guest Books Again				
	(1)	(2)	(3)	(4)	(5)
Review Submitted	0.093***	−0.031***	−0.013*		−0.001
	(0.001)	(0.007)	(0.007)		(0.010)
Positive Sentiment		−0.0003	−0.002		−0.012***
		(0.003)	(0.003)		(0.004)
Overall Rating		0.023***	0.017***	0.016***	0.017***
		(0.002)	(0.002)	(0.002)	(0.003)
Lowest Subrating		0.004***	0.005***	0.005***	0.004**
		(0.001)	(0.001)	(0.001)	(0.002)
Has Recommend				0.001	
				(0.012)	
Guest Recommends				0.026**	
				(0.012)	
Host Negative Sentiment			−0.082***	−0.082***	−0.081***
			(0.009)	(0.015)	(0.011)
Guest Experience Controls	Yes	Yes	Yes	Yes	Yes
Other Guest and Trip Char.	No	No	Yes	Yes	Yes
Listing FE	No	No	No	No	Yes
Only > 3 Stars	No	No	No	Yes	No
Observations	558,959	532,285	532,027	343,941	532,027

Table 1: The Informativeness of Reviews: Re-booking Rates

Re-booking rates are calculated from August 2014 to May 2015. The sample includes all trips in the incentivized review experiment. Experience controls are an indicator for whether the guest is new and the log of the number of prior trips plus one. Other controls include trip nights, guests, price per night, checkout date, guest age, guest region and listing region.

	Guest		Host	
	Control	Treatment	Control	Treatment
Reviews	0.257	0.426	0.626	0.632
Five Star	0.687	0.606	-	-
Recommends	0.963	0.954	0.986	0.985
High Likelihood to Recommend Airbnb	0.731	0.708	-	-
Overall Rating	4.599	4.488	-	-
All Sub-Ratings Five Star	0.458	0.389	0.805	0.795
Responds to Review	0.021	0.019	0.040	0.051
Private Feedback	0.432	0.439	0.275	0.273
Feedback to Airbnb	0.102	0.117	0.089	0.089
Mean Review Length (Sentences)	5.726	5.212	2.580	2.618
Negative Sentiment Given Not-Recommend	0.809	0.764	0.948	0.939
Text Classified Positive	0.814	0.874	0.806	0.838
Median Private Feedback Length (Characters)	131	126	96	95
First Reviewer	0.072	0.168	0.599	0.570
Time to Review (Days)	18.420	13.709	5.715	5.864
Time Between Reviews (Hours)	292.393	215.487	-	-
Num. Obs.	15470	15759	15470	15759

Table 2: Summary Statistics: Incentivized Review Experiment

Experiment:	**Coupon**	**Coupon**	**Sim. Reveal**	**Sim. Reveal**	**Coupon (Imputed)**
Sample:	Experimental Sample	No Prior Reviews	All Listings	No Prior Reviews	No Prior Reviews
Adjustment:		Effect on Distribution			If Everyone Reviewed
Specification:	(1)	(2)	(3)	(4)	(5)
Reviewed	0.166***	0.064***	0.018***	0.008	0.323
Five Star	-0.128***	-0.024***	-0.015***	-0.010*	-0.060
Recommends	-0.012	-0.004	-0.001	-0.001	-0.011
Neg. Sentiment	0.071**	0.008**	0.020***	0.028***	0.012

Table 3: Magnitudes of Experimental Treatment Effects

Columns (1), (3), and (4) display treatment effects in a linear probability model where the dependent variable is listed in the first column. The sample in column (1) only applies to observations in the coupon experiment that were also in the treatment of the simultaneous reveal experiment. Column (2) adjusts the treatment effects in column (1) to account for the fact that only guests who had not reviewed within 9 days were eligible for the coupon experiment. Therefore, the treatment effect in column (2) can be interpreted as the effect of the coupon experiment on average outcomes for all trips to non-reviewed listings. Column (4) presents the effects of the simultaneous reveal experiment for just the subsample of trips where the listing had no reviews. Column (5) displays predicted effects on reviews if everyone reviewed. Note that since this column is imputed, we do not compute associated standard errors. The regressions predicting ratings and sentiment are conditional on a submitted review. The controls are added to the regression are described in the text. *p<0.10, ** p<0.05, *** p<0.01

	(1)	(2)	(3)	(4)	(5)
Treatment	−0.082***	−0.081***	−0.116**	−0.077***	−0.088***
	(0.010)	(0.009)	(0.049)	(0.009)	(0.017)
Guest Lenient			0.156***		
			(0.057)		
Treatment * Guest Lenient			0.055		
			(0.075)		
Host Rev. First					0.073***
					(0.017)
Treatment * Host Rev. First					0.032
					(0.021)
Guest Characteristics	No	Yes	Yes	Yes	Yes
Listing Characteristics	No	No	No	Yes	Yes
Observations	10,626	10,626	584	10,626	10,626

Table 4: Effect of Coupon Treatment on Five Star Ratings

The table displays results of a regression predicting whether a guest submitted a five star rating in their review. "Treatment" refers to an email that offers the guest a coupon to leave a review. "Guest Lenient" is a an indicator variable for whether the guest previously gave higher than median ratings, as determined by a guest specific fixed effect in a regression on prior reviews. Guest controls include whether the guest is a host, region of origin, age, gender, nights of trip, number of guests, and checkout date. Listing controls include whether the host is multi-listing host, price, room type of the listing, and listing region. *p<0.10, ** p<0.05, *** p<0.01

	Guest Has Subequent Booking			
	(1)	(2)	(3)	(4)
Treatment	0.002	−0.019***	−0.012*	−0.012*
	(0.006)	(0.007)	(0.007)	(0.007)
Review Submitted		0.098***	0.066***	0.065***
		(0.009)	(0.009)	(0.009)
Treatment * Review Submitted		0.008	0.008	0.008
		(0.012)	(0.012)	(0.012)
Guest Characteristics	No	No	Yes	Yes
Listing Characteristics	No	No	No	Yes
Observations	29,481	29,481	29,481	29,481

Table 5: Selection into Reviewing: Guest Re-booking Rates

The table displays estimates from linear probability models. 'Guest Has Subsequent Booking' is defined as having a booking between September 2014 and May 2015. Guest controls include whether the guest is a host, region of origin, age, gender, nights of trip, number of guests, and checkout date. Listing controls include multi-listing host, price, room type, and region. p<0.10, ** p<0.05, *** p<0.01

	Guest		Host	
	Control	Treatment	Control	Treatment
Reviews	0.671	0.690	0.715	0.787
Five Star	0.741	0.726	-	-
Recommends	0.975	0.974	0.989	0.990
High Likelihood to Recommend Airbnb	0.766	0.759	-	-
Overall Rating	4.675	4.661	-	-
All Sub-Ratings Five Star	0.500	0.485	0.854	0.840
Responds to Review	0.025	0.066	0.067	0.097
Private Feedback	0.496	0.567	0.318	0.317
Feedback to Airbnb	0.106	0.109	0.068	0.072
Mean Review Length (Sentences)	5.393	5.454	2.926	2.915
Text Classified Positive	0.838	0.819	0.966	0.964
Negative Sentiment Given Not-Recommend	0.742	0.799	0.744	0.753
Median Private Feedback Length (Characters)	131	129	101	88
First Reviewer	0.350	0.340	0.491	0.518
Time to Review (Days)	4.284	3.897	3.667	3.430
Time Between Reviews (Hours)	63.680	47.478	-	-
Num. Obs.	60743	61018	60743	61018

Table 6: Summary Statistics: Simultaneous Reveal Experiment

	Host Sentiment	
	Negative	Positive
Low Rating	0.015	0.205
High Rating	0.014	0.766

(a) Control

	Host Sentiment	
	Negative	Positive
Low Rating	0.013	0.232
High Rating	0.018	0.737

(b) Treatment

Table 7: Cross Tabulation of Outcomes: Simultaneous Reveal Experiment

This table presents a cross-tabulation of review outcomes in the control and treatment groups of the simultaneous reveal experiment. ‘High Rating’ occurs when a guest submits a 5 star rating and ‘Low Rating’ occurs when the guests submits a lower than 5 star rating.

	Does Not Recommend	Overall Rating < 5	Negative Sentiment
	(1)	(2)	(3)
Treatment	0.001	0.028***	0.021***
	(0.002)	(0.005)	(0.004)
Host Negative Sentiment	0.560***	0.421***	0.439***
	(0.111)	(0.127)	(0.133)
Host Non-Recommend	0.123*	0.137	0.262**
	(0.073)	(0.102)	(0.104)
Treatment * Host Negative Sentiment	−0.333**	−0.314*	−0.219
	(0.132)	(0.175)	(0.181)
Treatment * Host Non-Recommend	−0.096	0.081	−0.093
	(0.085)	(0.145)	(0.146)
Guest, Trip, and Listing Char.	Yes	Yes	Yes
Observations	25,379	25,379	23,319

Table 8: Retaliation and Induced Reciprocity - Guest

This table presents the results of a linear regression predicting guest review behavior as a function of the simultaneous reveal treatment and host reviews, conditional on the host submitting a first review. There are fewer observations in column (3) because there was either too little associated text or it was in a different language. * p<0.10, ** p<0.05, *** p<0.01

	Rating > 3		
	(1)	(2)	(3)
Entire Property	-0.011^{***}	-0.013^{***}	
	(0.001)	(0.002)	
Multi-Listing Host	-0.045^{***}	-0.045^{***}	-0.032^{***}
	(0.001)	(0.002)	(0.003)
Recommended	0.758^{***}	0.742^{***}	0.691^{***}
	(0.001)	(0.001)	(0.002)
Multi-Listing * Recommended	0.031^{***}	0.030^{***}	0.030^{***}
	(0.001)	(0.002)	(0.002)
Entire Prop. * Recommended	0.014^{***}	0.015^{***}	0.010^{***}
	(0.001)	(0.002)	(0.002)
Guest FE	No	Yes	Yes
Market FE	Yes	Yes	No
Listing FE	No	No	Yes
Observations	2,274,159	2,274,159	2,274,159

Table 9: Socially Induced Reciprocity - Star Rating

The outcome in the above regression is whether the guest's star rating is greater than 3. The estimation is done on all trips between 2012 and 2014 for a 50% sample of guests. Controls for reviews, date, nights, customer support tickets, guest bookings, price, EPP, and listing capacity were included. *p<0.10, ** p<0.05, *** p<0.01

	Measure of Bias:	
Counterfactual:	B_{avg} Average	B_{neg} % Negative Missing
Baseline	1.32	61.24
Simultaneous Reveal	1.29	59.23
Simultaneous Reveal + No Social Reciprocity	1.14	56.65
Simultaneous Reveal + No Social Reciprocity + No Sorting	0.04	31.79
Above + Everyone Reviews	0.04	1.12

Table 10: Size of Bias
(Guest does not recommend listing but submits five star rating.)

The above table displays two measures of bias under five scenarios. The mis-reporting used to calculate bias occurs when the guest does not recommend the listing but omits any negative review text. B_{avg} is the difference between the average rate of negative sentiment for reviews (where the guest does not recommend), and the overall rate of trips where the guest has a negative experience. B_{neg} is share of all stays where a negative experience was not reported.

A Predictors of Review Rates

Table AIX displays the results of a linear probability regression that predicts whether a guest reviews as a function of guest, listing, and trip characteristics. Column 2 adds the city fixed effects in addition to the other variables. If worse experiences result in lower review rates, then worse listings should be less likely to receive a review. The regression shows that listings with lower ratings and lower historical review rates per trip have a lower chance of being reviewed. For example, a listing with an average review rating of four stars is .68 percentage points less likely to be reviewed than a listing with an average rating of five stars. Furthermore, trips where the guest calls customer service are associated with an 11% lower review rate.

Guest characteristics also influence the probability that a review is submitted. New guests and guests who found Airbnb through online marketing are less likely to leave reviews after a trip. This might be due to one of several explanations. First, experienced users who found Airbnb through their friends may be more committed to the Airbnb ecosystem and might feel more of an obligation to review. On the other hand, new users and users acquired through online marketing might have less of an expectation to use Airbnb again. Furthermore, these users might have worse experiences on average, either because they picked a bad listing due to inexperience or because they had flawed expectations about using Airbnb.

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 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".[14] Each review is transformed into a vector for which each column represents the presence of a word or phrase (up to 3 words), 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').

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

C Experimental Validity

This section documents that both experimental designs in this paper are valid. Table AII displays the balance of observable characteristics in the experiments. Turning first to the incentivized review experiment, the rate of assignment to the treatment in the data is not statistically different from 50%. Furthermore, there is no statistically significant difference in guest characteristics (experience, origin, tenure) or host characteristics (experience, origin, room type). Therefore, the experimental design is valid.

Similarly, there is no statistically significant difference in characteristics between the treatment and control guest in the simultaneous reveal experiment, However, there is a .3% difference between the number of observations in the treatment and control groups. This difference has a p-value of .073, making it barely significant according to commonly used decision rules. We do not know why this result occurs, although given that we make 14 comparisons in the table, it is likely that at least one of them will have a low p-value.

D Scaling Our Estimates

We use the following equation to adjust the experimental treatment effects to represent the overall effect on ratings for listings with 0 reviews.

$$e_m = \frac{s_{\leq 9} r_{m,\leq 9} + (s_{ctr} + t_{rev})(r_{m,ctr} + t_m)}{s_{\leq 9} + s_{ctr} + t_{rev}} - \frac{s_{\leq 9} r_{m,\leq 9} + s_{ctr} r_{m,ctr}}{s_{\leq 9} + s_{ctr}} \tag{11}$$

where e_m is the adjusted treatment effect for metric m, s refers to the share of trips in each group, t_m refers to the experimental treatment effect, and r_m refers to the mean value of a review metric, m. "≤ 9" refers to the sample of trips where the guest reviews within 9 days, "ctr" refers to the control group, and "rev" refers to the review rate.

We use the following equation to rescale the parameter governing the share of negative experiences from just those trips to non-reviewed listings to all trips. This calculation is used in Section

9.

$$\hat{g} = s_{0,ir,sr}\frac{s_{all,sr}}{s_{0,c,sr}} \tag{12}$$

where $s_{0,ir,sr}$ is the share of recommendations in the incentivized review (ir) and simultaneous reveal (sr) treatments, $s_{0,c,sr}$ is the share of recommendations in the ir control and sr treatment, and $s_{all,sr}$ is the share of positive reviews in the entire sr treatment.

E The Long-Run Evolution of Ratings

In this section, we document the distribution of ratings on Airbnb and how it changes over time. The reason we conduct this exercise is that our experiments may have accentuated effects due to reviewer ignorance about the review system changes. If knowing about the review system was important, then we would expect the average ratings to drop over time as people learned that they no longer need to fear retaliation because of simultaneous reveal reviews. Figure A3 displays the long-run trends in ratings as a share of all reviews for a set of experienced users, who should be aware of the workings of the review system. There are two key 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. However, 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.

F Additional Results on Strategic Reciprocity

In this section, we discuss results regarding strategic reciprocity for hosts who review second and guests who review first. Table AIV displays estimates for two outcomes: whether the host does not recommend and whether the host uses negative sentiment. For all specifications, the coefficient on the treatment is small and insignificant. Therefore, there is no evidence of induced reciprocity by positive guest reviews. However, there is evidence of retaliation in all specifications. Specifications (1) and (2) show that a low rating (< 5 stars) by a guest in the control is associated with a 27 percentage points lower recommendation rate and a 32 percentage point lower negative sentiment rate (defined across all host reviews regardless of the host's recommendation). The interaction with the treatment reduces the size of this effect almost completely. In specifications (3) and (4), we look at three types of initial guest feedback: recommendations, ratings, and negative sentiment conditional on not recommending the host. The predominant effect on host behavior across these three variables is the guest text. Guests' negative text increases hosts' use of negative text by 30 percentage points, while the coefficients corresponding to guests' ratings are relatively lower across specifications. This larger response to text is expected because text is always seen by the host whereas the rating is averaged across all prior guests and rounded. Therefore, hosts may not be able to observe and retaliate against a low rating that is submitted by a guest.

Guest's fear of retaliation can be seen in the effect of the treatment on private feedback to hosts. Guests have the ability to leave suggestions for a host to improve the listings. Private feedback cannot hurt the host, but it may still trigger retaliation. Table AVII displays the effect of the treatment on whether a guest leaves a suggestion. Column (1) shows that the overall effect of the treatment is 6.4 percentage points (se = .003), suggesting that guests are indeed motivated by fear of retaliation to not send private feedback to hosts. Columns (2) and (3) test whether this effect is driven by particular types of trips by interacting the treatment indicator with indicators for guests' recommendations and ratings. The effect of the treatment is more positive for guests who recommend the host but don't submit a five star rating. Therefore, the treatment allows guests who have good, but not great, experiences to offer suggestions to the host without a fear of retaliation.

G Additional Tables

	Guest Contacted Customer Support				
	(1)	(2)	(3)	(4)	(5)
Review Submitted	-0.008^{***}	0.096^{***}	0.094^{***}		0.075^{***}
	(0.0003)	(0.002)	(0.002)		(0.002)
Positive Sentiment		0.008^{***}	0.007^{***}		0.006^{***}
		(0.001)	(0.001)		(0.001)
Overall Rating		-0.021^{***}	-0.020^{***}	-0.001^{**}	-0.016^{***}
		(0.001)	(0.001)	(0.0004)	(0.001)
Lowest Subrating		-0.003^{***}	-0.003^{***}	-0.003^{***}	-0.003^{***}
		(0.0003)	(0.0003)	(0.0003)	(0.0004)
Has Recommend				0.013^{***}	
				(0.002)	
Guest Recommends				-0.014^{***}	
				(0.002)	
Guest Experience Controls	Yes	Yes	Yes	Yes	Yes
Other Guest and Trip Char.	No	No	Yes	Yes	Yes
Listing FE	No	No	No	No	Yes
Only $>$ 3 Stars	No	No	No	Yes	No
Observations	558,960	532,286	532,027	343,941	532,027

Table AI: The Informativeness of Reviews: Customer Support

Re-booking rates are calculated from August 2014 to May 2015. The sample includes all trips in the incentivized review experiment. Experience controls are an indicator for whether the guest is new and the log of the number of prior trips plus one. Other controls include trip nights, guests, price per night, checkout date, guest age, guest region and listing region.

Variable	Experiment	Difference	Mean Treatment	Mean Control	P-Value	Stars
Experienced Guest	Simultaneous Reveal	-0.001	0.557	0.558	0.702	
US Guest	Simultaneous Reveal	-0.001	0.282	0.283	0.761	
Prev. Host Bookings	Simultaneous Reveal	-0.162	14.875	15.037	0.272	
US Host	Simultaneous Reveal	0.001	0.263	0.262	0.801	
Entire Property	Simultaneous Reveal	-0.001	0.671	0.671	0.824	
Reviewed Listing	Simultaneous Reveal	-0.003	0.764	0.767	0.167	
Observations	Simultaneous Reveal	0.001			0.431	
Experienced Guest	Incentivized Review	-0.010	0.498	0.508	0.066	*
US Guest	Incentivized Review	0.001	0.228	0.227	0.859	
Prev. Host Bookings	Incentivized Review	-0.008	0.135	0.143	0.134	
US Host	Incentivized Review	0.0002	0.199	0.199	0.973	
Entire Property	Incentivized Review	0.002	0.683	0.681	0.645	
Host Reviews Within 7 Days	Incentivized Review	-0.009	0.736	0.745	0.147	
Observations	Incentivized Review	0.005			0.102	

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. Note, the sample averages for the two experiments differ because only guests to non-reviewed listings who had not reviewed within 9 days were eligible for the incentivized review experiment. *p<0.10, ** p<0.05, *** p<0.01

	Reviews First		Neg. Sentiment First	Num. Negative Phrases First
	(1)	(2)	(3)	(4)
Treatment	0.027***	0.026***	0.001	0.010***
	(0.003)	(0.003)	(0.002)	(0.003)
Customer Support		−0.191***		
		(0.020)		
Non-Recommend			0.664***	
			(0.035)	
Negative Sentiment				0.492***
				(0.056)
Treat. * Customer Support		0.080***		
		(0.030)		
Treat. * Non-Recommend			0.072	
			(0.044)	
Treat. * Neg. Sentiment				0.214***
				(0.082)
Guest, Trip, and Listing Char.	Yes	Yes	Yes	Yes
Observations	121,380	121,380	42,143	42,143

Table AIII: Fear of Retaliation - Host

This table presents the results of a linear regression predicting host review behavior as a function of the simultaneous reveal treatment and metrics of transactional quality. ‘Customer Support’ is an indicator for whether the host contacted customer support, ‘Non-recommend’ is an indicator for whether the host anonymously recommended the guest, and ‘Negative Sentiment’ is an indicator for whether the host’s text was classified as negative. ‘Num. Negative Phrases First’ is a count of separate negative n-grams in the host’s review of the guest. *p<0.10, ** p<0.05, *** p<0.01

	Does Not Recommend	Negative Sentiment	
	(1)	(2)	(3)
Treatment	-0.0003	0.008^{***}	0.007^{**}
	(0.001)	(0.002)	(0.003)
Non-Recommend	0.175^{**}	0.082	0.126
	(0.082)	(0.070)	(0.105)
Neg. Text and Non-Recommend	0.293^{***}	0.413^{***}	0.294^{**}
	(0.092)	(0.083)	(0.120)
<5 Rating			0.043^{***}
			(0.008)
Treatment * Non-Recommend	-0.155^{*}	-0.110	-0.158
	(0.086)	(0.070)	(0.105)
Treatment * Neg. Text and Non-Recommend	-0.213^{**}	-0.255^{***}	-0.148
	(0.098)	(0.088)	(0.125)
Treatment * <5 Rating			-0.022^{**}
			(0.010)
Guest, Trip, and Listing Char.	Yes	Yes	Yes
Observations	19,729	17,145	10,692

Table AIV: Retaliation and Induced Reciprocity - Host

The above regressions are estimated for the sample where the guest reviews first. Controls include whether there was a contact to customer support, the user type (new vs experienced, organic vs acquired by marketing), nights and number of guests of trip, whether the guest was a host, the age of the guest, the gender of the guest, whether the host is a property manger, the average review rating of the host, and the Effective Positive Percentage of the host. "Treatment" refers to the simultaneous reveal experiment. *p<0.10, ** p<0.05, *** p<0.01

	Measure of Bias:	
Counterfactual:	B_{avg} Average	B_{neg} % Negative Missing
Baseline	1.84	69.78
Simultaneous Reveal	1.69	65.98
Simultaneous Reveal + No Social Reciprocity	1.50	62.80
Simultaneous Reveal + No Social Reciprocity + No Sorting	0.56	41.47
Above + Everyone Reviews	0.56	15.15

Table AV: Size of Bias
(Guest does not recommend listing but omits negative text.)

The above table displays two measures of bias under five scenarios. The mis-reporting used to calculate bias occurs when the guest does not recommend the listing but omits any negative review text. B_{avg} is the difference between the average rate of negative sentiment for reviews (where the guest does not recommend), and the overall rate of trips where the guest has a negative experience. B_{neg} is share of all stays where a negative experience was not reported.

	Has Customer Service		Has Next Booking	
	(1)	(2)	(3)	(4)
Overall Rating	−0.021***	−0.021***	0.037***	0.031***
	(0.001)	(0.001)	(0.003)	(0.003)
Treatment	−0.006	−0.004	0.034	0.021
	(0.005)	(0.005)	(0.022)	(0.022)
Negative Text by Host				−0.064***
				(0.021)
Rating * Treatment	0.001	0.001	−0.008	−0.005
	(0.001)	(0.001)	(0.005)	(0.005)
Guest and Trip Controls	No	Yes	No	Yes
Observations	98,602	98,507	98,602	98,507
R^2	0.015	0.017	0.002	0.070

Table AVI: The Effect of Simultaneous Reveal on Rating Informativeness

This table presents the results of a linear regression predicting whether a guest contacted customer service during the trip and whether a guest books a trip between September 2014 and May 2015. Treatment refers to the Simultaneous Reaveal Treatment. Guest and trip controls include whether the guest was new, guest region, and listing region. *p<0.10, ** p<0.05, *** p<0.01

	Guest Left Private Suggestion for Host		
	(1)	(2)	(3)
Treatment	0.064***	0.046***	0.052***
	(0.003)	(0.004)	(0.007)
Customer Support	0.075***	0.082***	0.079***
	(0.019)	(0.019)	(0.019)
Guest Recommends		0.047***	0.052***
		(0.003)	(0.003)
Five Star Review			−0.074***
			(0.005)
Recommends * Treatment		0.022***	0.023***
		(0.004)	(0.004)
Five Star * Treatment			−0.012*
			(0.007)
Guest, Trip, and Listing Char.	Yes	Yes	Yes
Observations	82,623	82,623	82,623

Table AVII: Determinants of Private Feedback Increase

"Treatment" refers to the simultaneous reveal experiment. "Customer Support" refers to a guest initiated customer service complaint. Controls include the user type (new vs experienced, organic vs acquired by marketing), nights and number of guests of trip, whether the guest was a host, the age of the guest, the gender of the guest, whether the host is a property manger, and the five star review rate of the host. *p<0.10, ** p<0.05, *** p<0.01

	Rating > 3		
	(1)	(2)	(3)
Entire Property	0.0005	-0.046^{***}	-0.046^{***}
	(0.002)	(0.005)	(0.007)
High LTR			0.037^{***}
			(0.002)
Recommends		0.726^{***}	0.734^{***}
		(0.003)	(0.005)
Entire Prop. * Recommends		0.050^{***}	0.040^{***}
		(0.005)	(0.006)
Entire Prop. * High LTR			0.011^{***}
			(0.003)
Address FE	YES	YES	YES
Observations	232,899	205,085	112,783

Table AVIII: Socially Induced Reciprocity - Address Fixed Effects

The outcome in the above regression is whether the guest's star rating is greater than 3. The sample used is the set of trips to addresses the had multiple listing types, of which one had more than 1 bedroom, which took place between 2012 and 2014. "High LTR" occurs when the guest's likelihood to recommend is greater than 8 (out of 10). Controls for reviews, date, nights, guests, customer support, guest bookings, and log price were included. *p<0.10, ** p<0.05, *** p<0.01

	Reviewed	
Five Star Rate	0.106***	0.106***
	(0.008)	(0.008)
Past Booker	0.058***	
	(0.004)	
No Reviews	0.028**	0.028**
	(0.013)	(0.013)
No Trips	0.095***	0.096***
	(0.012)	(0.012)
Num. Trips	−0.0005***	−0.0005***
	(0.0001)	(0.0001)
Customer Service	−0.175***	−0.169***
	(0.020)	(0.020)
Entire Property	0.004	0.005
	(0.005)	(0.005)
Multi-Listing Host	−0.100***	−0.089***
	(0.007)	(0.007)
Log Price per Night	−0.011***	−0.012***
	(0.003)	(0.003)
Trip Characteristics	Yes	Yes
Market FE:	No	Yes
Observations	60,552	60,552
Note:	*p<0.1; **p<0.05; ***p<0.01	

Table AIX: Determinants of Guest Reviews

These regressions predict whether a guest submits a review conditional on the observed characteristics of the listing and trip. Only observations in the control group of the simultaneous reveal experiment are used for this estimation.

H Additional Figures

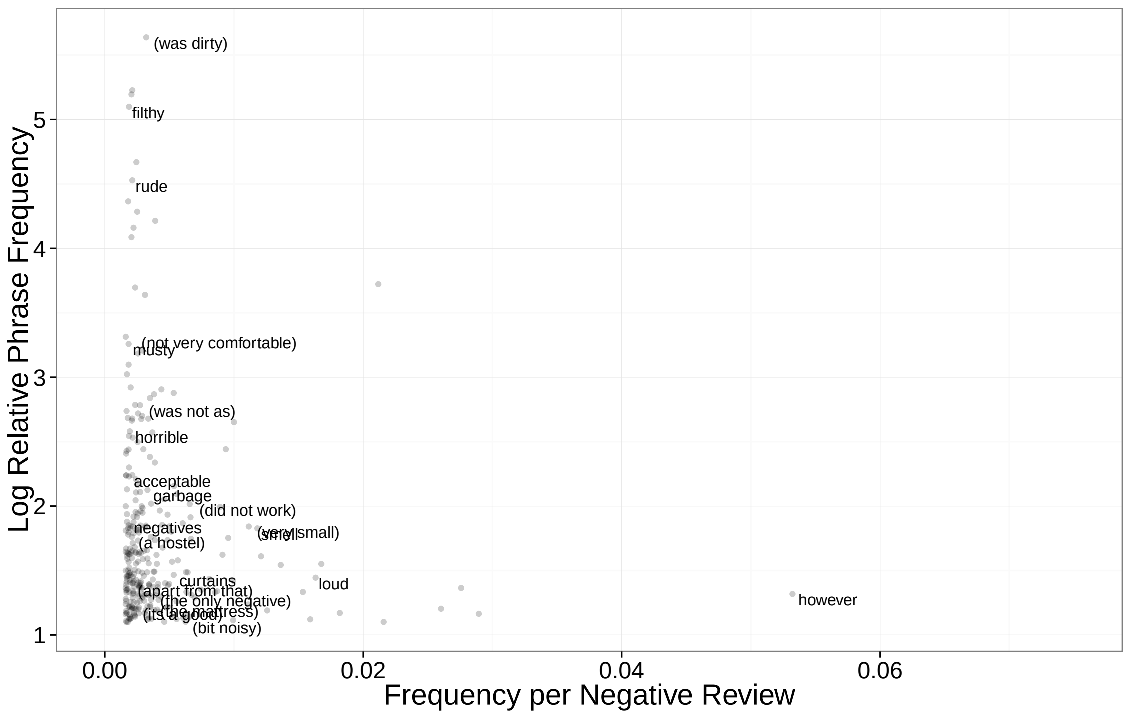


Figure A1: Distribution of negative phrases in guest reviews of listings.

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

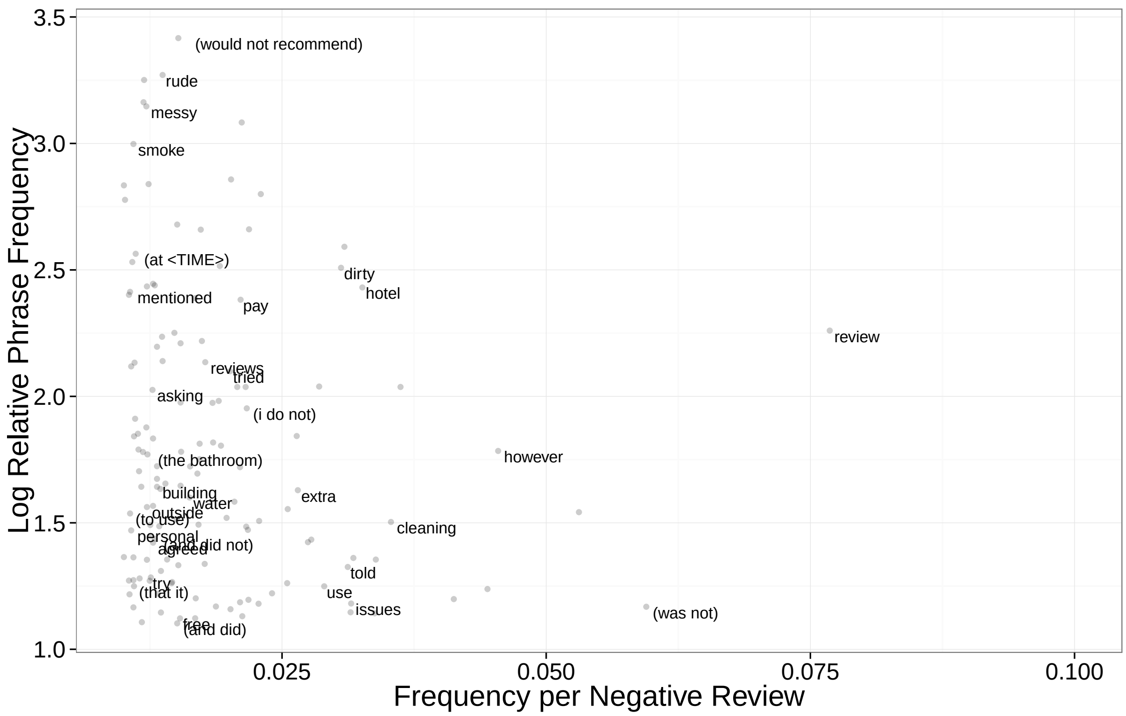


Figure A2: Distribution of negative phrases in host reviews of guests.

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

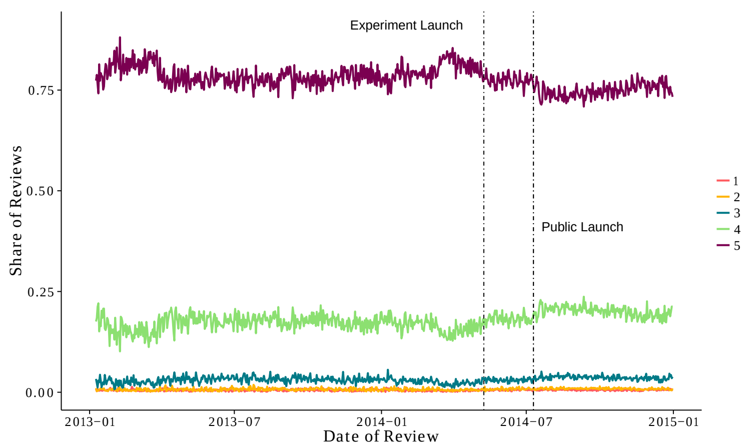


Figure A3: Ratings Over Time

This figure displays the temporal trends of review ratings over time. Because composition of guests and hosts varies with the growth of the platform, this figure is for 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.