Bias and Reciprocity in Online Reviews: Evidence From Field Experiments on Airbnb (Preliminary - Do not cite without permission)

Andrey Fradkin*[1], Elena Grewal†[2], David Holtz‡[1], and Matthew Pearson§[2]

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

December 26, 2015

Abstract

Reviews and other evaluations are used by consumers to decide what goods to buy and by firms to choose whom to trade with, hire, or promote. However, because potential reviewers are not compensated for submitting reviews and may have reasons to omit relevant information in their reviews, reviews may be biased. We use the setting of Airbnb to study the determinants of reviewing behavior, the extent to which reviews are biased, and whether changes in the design of reputation systems can reduce that bias. We find that reviews on Airbnb are typically positive and informative. 97% of guests privately and anonymously report having positive experiences and 74% of guests submit the maximum score, five stars. When guests do not recommend a listing, this is reflected in a lower than five star rating over 90% of the time. We use the results from two field experiments intended to reduce bias to document that non-reviewers tend to have worse experiences than reviewers and that strategic reviewing behavior occurred on the site, although the aggregate effect of the strategic behavior was relatively small. Lastly, we document the presence of socially induced reciprocity in reviews, by which more social trips result in lower reporting rates of negative experiences. We use a model to show that these mechanisms for bias decrease the rate of reviews of negative experiences by 1.28 percentage points relative to scenario where all transactions resulted in honest reviews.

We are grateful to Jon Levin, Liran Einav, Caroline Hoxby, Shane Greenstein, Ramesh Johari, Mike Luca, Chris Dellarocas, John Horton, Chiara Faronnato, Jeff Naecker, Fred Panier, and seminar participants at Microsoft, eBay, ACM EC'15, NBER Summer Institute, and the CODE Conference for comments. Note: The views expressed in this paper are solely the authors' and do not necessarily reflect the views of Airbnb Inc.

*Primary Author: fradkina@nber.org

†Primary Experiment Designer: elena.grewal@airbnb.com

‡dave.holtz@airbnb.com

§matthew.pearson@airbnb.com

1 Introduction

Reviews and other evaluations are used by consumers to decide what goods to buy and by firms to choose whom to trade with, hire, or promote. These reviews are especially important for online marketplaces (e.g. eBay, Amazon, Airbnb, and Etsy), where economic agents often interact with new trading partners who provide heterogeneous goods and services.[1] However, potential reviewers are not compensated for submitting reviews and may have reasons to omit relevant information when reviewing. Therefore, basic economic theory suggests that accurate reviews constitute a public good and are likely to be under-provided (Avery et al. (1999), Miller et al. (2005)). As a result, the distribution of evaluations for a given agent may not accurately represent the outcomes of that agent's previous transactions. The presence of this review bias may consequently reduce market efficiency. For example, it may cause agents to engage in suboptimal transactions (Horton (2014), Nosko and Tadelis (2015)) or to leave the platform altogether.

We study the determinants of reviewing behavior, the informational content of online reviews, and the effects of changes in the design of reputation systems. The setting of this paper is Airbnb, a large online marketplace for accommodations. Reputation is thought to be particularly important for transactions on Airbnb because guests and hosts interact in person, often in the primary home of the host.[2] As in many other marketplaces, reviews are predominantly positive.[3] Over 70% of the guests in our sample submit a five (out of a maximum of five) star rating for their hosts. These positive ratings may genuinely reflect positive experiences of guests or the may be the result of biases in the reputation system.

We use proprietary data from Airbnb to show that, although bias does exist, the high average rating is consistent with private and anonymous signals of transaction quality. Our empirical strategy relies on the fact that Airbnb's review system solicits anonymous recommendations which are never shown to the transaction partner or to future guests. Reviewers should have no reason to misreport their experience in this anonymous recommendation. We find that over 97% of reviewers recommend their counterparty and that when they do not recommend their counter-party, they leave a lower than five star ratings over 90% of the time. Furthermore, public review ratings predict customer service calls during a transaction and the rebooking rates of guests, indicating that these ratings are informative.

Although reviews are informative, bias may still exist in the review system. We describe a theoretical framework for measuring bias and use experiments as well as non-experimental analysis to study its magnitude. In our theoretical framework, there are two conditions under which reviews can be biased.[4] First, those that review an agent might differ systematically in their experiences from those that do not review an agent. Second, reviewers might not

[1] There is a large literature studying the effects of reputation scores on market outcomes. Pallais (2014) uses experiments to show that reviews affect demand for workers on Odesk. Cabral and Hortaçsu (2010) use panel data to show that reputation affects exit decisions by firms on eBay. Luca (2013) shows that Yelp reputation has especially large effects on non-chain restaurants.

[2] For example, Friedman (2014) wrote the following in the New York Times: "Airbnb's real innovation — a platform of 'trust' — where everyone could not only see everyone elses identity but also rate them as good, bad or indifferent hosts or guests. This meant everyone using the system would pretty quickly develop a relevant 'reputation' visible to everyone else in the system."

[3] For example, Horton (2014) shows that 91% of ratings on Odesk in 2014 were four or five (out of five) stars.

reveal their experiences in the public review. The extent of bias is a function of the utility of reviewing and the design of the reputation system. For example, reviewers who care about their own reputation may be afraid of retaliation by the counterparty and may consequently inflate their reviews. Changes in the reputation system that remove the possibility of retaliation would then make reviews more consistenet with buyer experiences and increase review rates. We show that the gains to a more informative review system are a function of the share of low quality sellers, the disutility from transacting with these sellers, and the rate at which transactions with these sellers are reported in reviews.

Our first field experiment, described in more detail in section 4, offers a $25 coupon as an incentive for guests to leave a review. The treatment group experiences a 6.4 percentage point increase in review rates and the share of trips that are rated five stars by 2.1 percentage points. However, conditional on reviewing, those in the treatment group were 7 percentage points less likely to submit a five star rating. This effect has two potential causes. First, those induced to review could have had a worse experience than those who would have reviewed regardless. Second, the monetary incentives from the coupon could have changed the propensity of reviewers to review positively. We show that selection is the primary channel of the effect since the coupon does not cause guests to change their reviewing style and the return rates of reviewers in the treatment are lower than the return rate of reviewers in the control.

The second condition for review bias occurs when reviewers do not reveal their experiences in the review. We show that this misrepresentation does occur in the data. For example, 6% of guests who anonymously answered that they would not recommend their host nonetheless submitted a public review with a five star rating. One possible reason for this misrepresentation is strategic behavior on behalf of reviewers. For example, Cabral and Hortaçsu (2010) and Saeedi et al. (2015) show that when eBay had a two sided review system, over 20% of negative buyer reviews were followed by negative seller reviews, interpreted by the authors as retaliatory. Bolton et al. (2012) use an innovative laboratory experiment to study the effects of a simultaneous reveal system in which reviews are hidden until both parties submit a review ("simultaneous reveal"). They show that such a system reduces retaliation and makes markets more efficient.

We document the first experimental test of such a simultaneous reveal mechanism in an actual online marketplace and use it to test whether the effects of the experiment in the lab are applicable in the field. The experiment we study was conducted by Airbnb to determine the effect of the change in mechanisms as a part of its product design process (following the experiment, Airbnb released the simultaneous review treatment to all users). The treatment increased review rates by guests while decreasing the share of five star reviews by 1.6 percentage points. On the host side, the treatment increased review rates by 7 percentage points but does not affect recommendation rates. We show that strategic motives affected reviewing behavior in the control group by demonstrating that the relationship between the first reviewer's review and the second reviewer's review changes due to the experiment. Our results differ from the laboratory results in Bolton et al. (2012) in two

[4] There is also considerable evidence about fake promotional reviews, which occur when firms post reviews either promoting themselves or disparaging competitors (see (Mayzlin et al., 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.

ways. First, our experimental treatment effect on five star ratings of -1.5 percentage points is smaller than their -7.7 percentage point effect found in a laboratory setting. Second, our simultaneous reveal mechanism experiment increases review rates while the same mechanism caused reductions in review rates in the lab.

One reason that the simultaneous reveal treatment has a relatively small effect may be that there are non-strategic reasons why reviewers omit information. We show that even in the simultaneous reveal treatment group, 7% of non-recommendations are accompanied by five star ratings and 13% are accompanied by positive review text. In section 6 we use non-experimental evidence to study this mismatch. We find that mismatch between public and private ratings in the cross-section is predicted by property type (entire home or a room in a home) and host type (multi-listing host or casual host). We use two distinct identification strategies to show that the coefficients on these characteristics likely represent causal effects.

First, we compare guest reviewing behavior in cases when a given host sometimes rents out her entire place and other times just a room. We find that guests to the private room are more likely to submit a four or five star rating when they do not recommend the listing. Second, we consider cases when a host who was once a casual host became a multi-listing host. We find that the rate of mismatch decreases when the host becomes a multi-listing host.

We hypothesize that these effects occur because buyers and sellers sometimes have a social relationship. For example, guests who rent a room within a property may talk to their hosts in person. Alternatively, guests may feel more empathy for casual hosts rather than multi-listing hosts, who may communicate in a more transactional manner. Social communication can lead reviewers to omit negative comments due to two reasons. First, conversation can cause buyers and sellers to feel empathy towards each other (Andreoni and Rao (2011)). This may cause buyers to assume that any problem that occurs during the trip is inadvertent and not actually the fault of the seller. Second, social interaction may cause buyers to feel an obligation towards sellers because those sellers offered a service and were "nice" (Malmendier and Schmidt, 2012). This obligation can lead buyers to omit negative feedback because it would hurt the seller or because it would be awkward.[5]

Lastly, we conduct a quantitative exercise to measure the magnitude of bias in online reviews. We define bias as occurring when a negative experience does not result in a negative public review. We first show that private and public review mismatch predicts guests leaving the platform and the presence of customer service transactions during the transaction. We calculate that bias decreases the rate of reviews with negative text and a non-recommendation by .86 percentage points. This result is due to the fact that most guests respond that they would recommend their host. However, although the overall bias is small, when negative guest experiences do occur, they are not captured in the review text 70% of the time. We find that most of this effect is caused by sorting into reviewing and the fact that not everyone reviews. This suggests that inducing additional reviews and displaying data on non-reviews can increase market efficiency.

[5] Airbnb has conducted surveys of guests who did not submit a review asking why they did not submit one. Typical responses include: "Our host made us feel very welcome and the accommodation was very nice so we didn't want to have any bad feelings". "I also assume that if they can do anything about it they will, and didn't want that feedback to mar their reputation!"

Empirical Context and Related Literature:

Our empirical strategy has at least three advantages over the prior literature on bias in online reviews. First, we conduct two large field experiments that vary the incentives of reviewers on Airbnb. This allows us to credibly identify the causal effects of changes to review systems. Second, we use proprietary data which is observed by Airbnb but not by market participants. This gives us two pieces of information, transactions and private review information, which are typically not used by prior studies. We can use this data to study selection into reviewing and differences between the publicly submitted review and the privately reported quality of a person's experiences. Lastly, Airbnb (along with Uber, Taskrabbit, Postmates, and others) is a part of a new sector, often referred to as the "Sharing Economy", which facilitates the exchange of local services and underutilized assets between buyers and semi-professional sellers. There has been relatively little empirical work on this sector.[6]

Other evidence about Airbnb reviews comes from comparisons with hotel reviews. Zervas et al. (2015) compare the distribution of reviews for the same property on both TripAdvisor and Airbnb and shows that ratings on Expedia are lower than those on Airbnb by an average of at least .7 stars. More generally, the rate of five star reviews is 31% on TripAdvisor and 44% on Expedia (Mayzlin et al. (2014)) compared to 75% on Airbnb. This difference in ratings has led some to conclude that two-sided review systems induce bias in ratings. Our analysis suggests that the five star rate on Airbnb would be substantially higher than 44% even if the three forms of bias that we consider are removed.

There are other potential explanations for the observed differences in ratings distributions between platforms. For example, a much lower share of bookers submit a review on Expedia than on Airbnb.[7] This may lead reviews on Expedia to be negatively biased if only guests with extreme experiences submit reviews. Alternatively, guests on Airbnb and guests of hotels may have different expectations when they book a listing. A particular listing may justifiably receive a five star rating if it delivered the experience that an Airbnb guest was looking for at the transaction price, even if an Expedia guest would not have been satisfied.[8]

Numerous studies have proposed theoretical reasons why bias may occur but most of the evidence on the importance of these theoretical concerns is observational or conducted in a laboratory setting. For example, Dellarocas and Wood (2007) use observational data from eBay to estimate a model of reviewing behavior.[9] They show that buyers and sellers with mediocre experiences review fewer than 3 percent of the time. Although our experimental results confirm that mediocre users are less likely to review, the selection is less severe. Nosko and Tadelis (2015) show that eBay's search algorithms create better matches when they account for review bias using a sellers Effective Positive Percentage (EPP), the ratio of positive reviews to transactions (rather than total reviews). We provide the first causal

[6] Although see recent contributions by Fradkin (2014) about Airbnb and Cullen and Farronato (2015) about Taskrabbit.

[7] A rough estimate of review rates on Expedia can be derived as follows. Expedia had approximately $30 billion in bookings in 2012 and approximately 1 million reviews (`http://content26.com/blog/expedias-emily-pearce-user-reviews-rule-the-roost/`). If trips have an average price of $1000 then the review rates on Expedia are around 3%. In comparison, review rates on Airbnb are over 70%.

[8] Below, we list three other reasons why the distribution of reviews on Airbnb and hotel review sites may differ. One, the price a given listing charges on the two sites may be different. Two, TripAdvisor in particular is prone to fake reviews which tend to deflate overall ratings (Mayzlin et al. (2014)). Three, low rated listings may be filtered out at different rates between various platforms.

evidence that buyers who dont review have worse experiences and, by doing so, provide support for using the EPP metric.

Our coupon intervention reduced bias, but, because coupons are expensive and prone to manipulation, this intervention is not scalable. Li and Xiao (2014) propose an alternative way to induce reviews by allowing sellers to offer guaranteed rebates to buyers who leave a review. However, Cabral and Li (2014) show that rebates actually induce reciprocity in buyers and increase the bias in reviews.

There are other potential problems with review systems which we do not study. Reviews may be too coarse if many types of experiences are considered by guests to be worthy of five stars. Another potential problem is that reviewers may react in response to existing reviews (e.g. Moe and Schweidel (2011) and Nagle and Riedl (2014)). Because reviewers on Airbnb typically enter the review flow through an email or notification, they are unlikely to be reading prior reviews when choosing to submit a review and would have to remember the reviews they read when booking for this channel to be important. Lastly, even in an unbiased review system, cognitive constraints may prevent agent from using all of the available review information to make decisions, creating predictably suboptimal transactions.

There are several parallels between social influence in reviewing behavior and social influence in giving experiments. Bohnet and Frey (1999) use laboratory experiment to show that giving decreases with social distance and (Sally (1995)) shows that giving increases with non-binding communication. Anonymity is another important factor in giving behavior. For example, Hoffman et al. (1994) and Hoffman et al. (1996) find that giving decreases with more anonymity and increases with language suggesting sharing. Since transactions on Airbnb are frequently in person, involve social communication, and are branded as sharing, they represent a real world analogue to the above experiments.

Similarly, Malmendier et al. (2014), Lazear et al. (2012), and DellaVigna et al. (2012) find that when given the choice, many subjects opt-out of giving games. When subjects that opt-out are induced to participate through monetary incentives, they give less than subjects that opt-in even without a payment. We find the same effect with regards to reviews — when those that opt-out of reviewing are paid to review, they leave lower ratings. Our results are therefore consistent with models in which leaving a positive review is an act of giving from the reviewer to the reviewed.

2 Setting, Descriptive Statistics, and the Informativeness of Ratings

Airbnb describes itself as a trusted community marketplace for people to list, discover, and book unique accommodations around the world. Since 2008, Airbnb has accommodated over 30 million guests and has listed over one million listings. 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

[9] Also see Dai et al. (2012) for an interesting use of a structural model to infer restaurant quality and the determinants of reviewing behavior using the sequence of observed Yelp reviews.

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 2 pages asking public, private, and anonymous questions (shown in Figure 1). Guests are first asked to leave feedback consisting of publicly shown text, a one to five star rating,[10] 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 communicativeness of the host, the cleanliness of the listing, the location 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 3 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 communicativeness 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. Fradkin (2014) shows that, conditional on observable characteristics, reviewed guests experience lower rejection rates by potential hosts. Finally, the host can provide private text feedback about the quality of their hosting experience to the guest and to Airbnb.

2.1 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.[11] The summary statistics for these trips are shown in Table 1. Turning first to review rates, 67% of trips result in a guest review and 72% result in a host review. Furthermore, 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.

[10] 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.

[11] 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.

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 misstate the truth. Figure 2 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. In the next section, we show that mismatch between private and public ratings predicts future guest outcomes, further confirming that at least some of these cases are not mistakes.[12]

The text of a review is the most public aspect of the information collected by the review system because the text of a review is permanently associated with an individual reviewer. There is also evidence that review text influences consumer decisions even when star ratings are present (Archak et al. (2011)). Review text can contain a variety of information about the quality of a transaction and the characteristics of a product. 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 two approaches to measure sentiment. The first and preferred strategy uses 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. This approach is described in greater detail in Appendix A.

The most important choice in this procedure is what data to use to "train" (estimate) the model. Our training sample for guest reviews of hosts consists of reviews with five stars, which are labeled as "positive", and reviews with one or two stars, which are labeled as "negative". Our training sample for host reviews of guests uses either a non-recommendation or a sub-rating that is lower that 4 stars as a negative label and a recommendation with all sub-ratings as five stars as a positive label. After training the models, we apply them to predict the sentiment in the set of reviews we study for this paper. Both the guest and host review samples are taken from the period before the experiments so that the model training is not affected by the experimental we study. As an alternative classification strategy, we code whether a review had at least one negative word or phrase. A word or phrase is considered negative if it appears three times as frequently in reviews with negative recommendations as reviews with five star ratings and recommendations. The word or phrase must also meet a minimum frequency threshold.

Phrases that commonly show up in negative reviews by guests concern cleanliness, smell, unsuitable furniture, noise, and sentiment (see Figure A1 for specific examples). In Figure 3 we show the share of reviews with negative text conditional on the rating. Over 90% of 1

[12]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.

and 2 star reviews are classified as negative and these reviews contain the most common negative phrases at 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 leave 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. In the rare cases when negative reviews by hosts do occur, they contain phrases concerning not recommending the guest, personal communication, money, cleanliness, and damage (see Figure A2 for examples).

2.2 Are Reviews Informative?

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 review information is indeed informative by showing that higher ratings predict future re-booking rates by guests and the presence of customer service tickets during the trip.[13]

Table 2 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 those are particularly informative about future re-booking rates. In order to improve statistical power, the sample includes all trips in the subsequent experiments. Column (1) shows a baseline specification that shows that guests who review are 9.3 percentage points 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 2.3 percentage point 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.

So far, we've only used publicly visible review information in predicting ratings. In column (4) of Table 2 we focus on cases where the public rating is high (greater than 3 stars)

[13]In order to conserve space, we relegate the results on customer service to Appendix Table AII

and look at the informativeness of the private and anonymous recommendations. Conditional on a star rating, a guest non-recommendation is associated with a 2.6 percentage point decrease in re-booking rates. Therefore, the recommendation contains additional information not captured in the star ratings.

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.

3 Theoretical Framework for Review Bias and Its Effects

In this section we describe a simple model of review bias and how reviewing behavior affects market efficiency. Suppose there is a marketplace that brings together buyers and sellers. 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 2 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 $\geq$ 0 and their marginal cost is 0. Sellers do not know their type in the first period.

On the demand side, there are K $>$ 1 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 $\underline{u} \leq \frac{(1-\mu)u_l + .5\mu u_h}{1-.5\mu}$. 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:

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

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 being dishonest. In the case of an interaction with a low quality seller, buyers have to make a choice between misrepresenting their experience, telling the truth, 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. One argument against the importance of this observation is that if there were more periods, than all low sellers would eventually get a negative review and would be identified as low quality. In practice, review ratings are rounded to the nearest half a star and sometimes even good sellers get bad reviews (although this possibility is not included in the mode for expositional purposes). Therefore, buyers still face situations where multiple seller types have the same rating.

Observation 2: The platform knows more information that the buyers do not know about the likely quality of a seller.

Since high type sellers are more likely to be reviewed, 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. Second, platforms often observe private signals associated with a given transaction or review and can use that information to identify the quality of a seller. In our setting, Airbnb can see guests' anonymous recommendations and customer service calls.

To analyze the welfare implications of bias, we will change the notation slightly. Let r_p be equal to the probability that a buyer who transacts with an H seller leaves a positive review, let r_{lp} be the probability that a buyer who transacts with an L seller leaves a positive review, and let r_{ll} be the probability that a buyer who transacts with an L seller leaves a negative review. These probabilities are functions of the utility of review parameters (α_i, β_i, γ).

All sellers without a negative review transact because buyers' expected utility from the transaction is higher than their reservation utility. The welfare gains from having the marketplace (excluding the disutility from reviewing) are:

$$Welfare = \mu u_h + (1-\mu)(1-.5r_{ll})u_l - (1-.5(1-\mu)r_{ll})\underline{\mathrm{u}} \tag{2}$$

Now suppose that everyone reviewed and did so honestly. This corresponds to $r_p = 1$ and $r_{ll} = 1$. The difference in welfare between the scenario where everyone reviews honestly and the status quo is $.5(1-\mu)(\underline{\mathrm{u}} - u_L)(1 - r_{ll})$.

Therefore, the gain from having a better review system is a function both of the prevalence of bad actors (1 - μ), the probability guests submit honest negative reports about transactions with low quality listings, r_{ll}, and the disutility from transacting with a low quality seller compared to the utility from the outside option. This analysis justifies our focus on cases where a negative experience is not reported either due to a lack of review, which occurs with probability $1 - r_{ll} - r_{lp}$, or a misreported review, which occurs with probability r_{lp}. If we take the high rates of positive recommendations on Airbnb at face value, then $1-\mu$ is close to 0. In that case, an imperfect review system only causes large welfare losses on the platform when the utility from negative experiences, u_l, is very low relative to the outside option. This is much more likely to be the case for transactions in the sharing economy, then in purchases of physical goods, which can often be returned and typically do not cause disutility, even if they are bad.

Next, consider the effects of changing the parameters related to reviewing. Increasing β_i can induce additional buyers to review but it does not change their decision to report honestly or dishonestly. Because welfare depends only on negative reviews, the gains from increasing β_i come from inducing buyers who were previously not reviewing to honestly review. In our results, this parameter change corresponds to offering a coupon for buyers to review, where that coupon does not affect α_i.

Increasing α_i induces additional buyers to review positively and induces truth-tellers to misreport. The welfare change from increasing α_i comes from inducing dishonest reports because only r_{ll} matters for welfare. Tying this to our later empirical results, increasing socially induced reciprocity corresponds to an increase in α_i. Therefore, while socially induced reciprocity increases review rates for high quality sellers, it also increases misreporting for low quality sellers and therefore reduces market efficiency in this model.

4 The Incentivized Review Experiment

In this section we study the results of an experiment intended to induce additional reviews. In the experiment, which was conducted between April and July of 2014, all trips to non-reviewed listings for which the guest did not leave a review within 9 days were assigned to either a treatment group or a control group, each assigned with a 50% probability at a host level. Guests in the treatment group received an email offering a \$25 Airbnb coupon while guests in the control group received a normal reminder email (shown in Figure 4).

The treatment affected the probability of a review and consequently the probability of additional bookings for a listing. This resulted in more trips in the experimental sample to listings in the control group than listings in the treatment group. Therefore, we limit the analysis to the first trip to a listing in the experiment. Appendix C demonstrates that the randomization for this experiment is valid.

Table 3 displays the review related summary statistics of the treatment and control groups in this experiment. First, note that the 23% 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 12%. The left panel of figure 5 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 2 percentage point decrease in the recommendation and likelihood to recommend Airbnb rates.

Because only those guests who had not left a review within 9 days are eligible to be in the experiment, the estimated treatment effects do not represent changes to the overall distribution of ratings for non-reviewed listings. We use the following equation to adjust the experimental treatment effects to represent the overall effect on ratings for listings with 0 reviews.

$$e_m = \frac{s_{\leq 9} r_{m,\leq 9} + (s_{ctr} + t_{rev})(r_{m,ctr} + t_m)}{s_{\leq 9} + s_{ctr} + t_{rev}} - \frac{s_{\leq 9} r_{m,\leq 9} + s_{ctr} r_{m,ctr}}{s_{\leq 9} + s_{ctr}} \tag{3}$$

where e_m is the adjusted treatment effect for metric m, s refers to the share of trips in each group, t 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 t_{rev} refers to the treatment effect of the experiment on review rates.

Table 4 displays the baseline treatment effects (Column 1) and adjusted treatment effects (Column 2) 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).[14] The 17 percentage point treatment effect on review rates in the experiment drops to a 6.4 percentage point effect when scaled. Because of this scaling, the effect of the experiment is smaller on the overall distribution of reviews than on the distribution of reviews in the experiment. Another reason why there is a difference between columns (1) and (2) is that guests 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. In total, the experiment decreases the overall share of five star ratings by 2.4 percentage points and the share of reviews with recommendations by .8 percentage points.

4.1 Sorting or Crowding Out of Pro-Social Motives?

There are two 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 with different types of individuals, trips, or experiences compared to the control. Second, the presence of a monetary incentive may have changed the reviewing behavior of those that would have already left a review (e.g. Benabou and Tirole (2006)). We test the relative importance of these effects by looking at differences in ratings and rebooking rates across experiments.

First, we test whether the effect of the treatment on ratings persists after adding controls for a guest's prior review leniency and observed characteristics. If guests change their reviewing patterns due to the coupon, then we would expect that the monetary incentive had an effect in addition to sorting. Column (1) of Table 5 displays the baseline treatment effect of the experiment without any controls. Column (2) add controls for guest origin, experience, and trip characteristics. The effect of the treatment remains of the same magnitude, demonstrating that the treatment effects are not driven by selection on observable guest characteristics. Column (3) shows estimates 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 covariates, is greater than the median.[15] As expected, the coef-

[14] The effect of the coupon was larger in the treatment group of the simultaneous reveal experiment. We discuss this result in subsection 5.1.

ficient on the guest leniency term is positive, with more lenient guests leaving higher ratings. Adding this control does not diminish the effect of the experiment on ratings. Furthermore, the interaction between the treatment and guest judiciousness is not statistically significant. Therefore, the rating behavior of these guests, conditional on submitting a review, does not eliminate the baseline effect of the coupon. In column (4), we test whether more negative reviews are driven by listing composition. Adding controls for listing type, location, price, and number of non-reviewed stays increases the treatment effect to 7.7 percentage points. These results support the hypothesis that the coupon works mainly by inducing those with worse experiences, conditional on observables, to submit reviews.

An alternative way to test whether there is sorting is to look at guest rebooking rates, which should be correlated with the quality of guest experience. Table 6 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. Second, column (2) shows that there is a 1.9 percentage point difference in rebooking rates between the treatment and the control group and that reviewers are 9.8 percentage points more likely to rebook. Reviewers in the treatment group are 1.1 percentage points less likely to rebook, although this difference is not statistically significant. These magnitudes are reduced but not eliminated by including guest and listing characteristics in columns (3) and (4).

We interpret the evidence from these specification in the following manner. First, the fact that non-reviewers are less likely to return confirms there is selection into reviewing based on the quality of a guest's experience. Second, the fact that non-reviewers in the treatment have lower rebooking rates suggests that those induced to review by the coupon tended to i) have better experiences than those not induced to review or ii) have valued the coupon more because they were going to return to the site anyway.

In summary, our results suggest that those who do not review have systematically worse experiences than those that do review. Those induced to review by the experiment have systematically worse experiences than those who review in the control group. Furthermore, our results actually understate the extent of sorting, because those not induced to review have even lower re-booking rates than those induced to review. In order to extrapolate from the experiment to the total effect of sorting on the ratings distribution of non-reviewed listings, we need to make an assumption regarding the experiences of non-reviewers. As a conservative estimate of the bias due to sorting we assume that non-reviewers had the same experiences on average as reviewers in the treatment group of the experiment. Column (5) of Table 4 displays these imputed selection effect. Under this assumption, there would be a 6 percentage point lower five star review rate and a 6.8 percentage point higher rate of negative text.

5 The Simultaneous Reveal Experiment

In this section we study the effects of a change in Airbnb's review system intended to remove strategic retaliation and reciprocation of reviews. Prior to May 8, 2014, both guests and

[15]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.

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 retaliates or reciprocates the first review. Futhermore, 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 this strategic reciprocity 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 (shown in Figure 6). 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.

For this analysis we limit the data we use to the first trip 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 7 shows the summary statistics for the treatment and control groups in the "simultaneous reveal" experiment. The treatment increases review rates for guests by 2 percentage points and for hosts by 7 percentage points. The rate of five star reviews by guests decreases by 1.6 percentage points, while the recommendation rate decreases by .4 percentage points. Furthermore, the drop in positive text by guests of 1.5 percentage points mirrors the drop in five star reviews. This suggests that the fear of retaliation had a similar effect on both the averaged ratings and the text in which the reviewer was identifiable. The treatment induced a 6.4 percentage points higher rate of guest suggestions to hosts. (see Table AVI). This increase was present even when conditioning on guest recommendations and star ratings. The relatively larger increase in private feedback suggests that without the fear of retaliation, guests felt they could speak more freely to the hosts about problems with the listing.[16]

Columns (3) and (4) of Table 4 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. Of note is that although the experiment has statistically significant effects on reviewing behavior, they are generally smaller than the effects of the coupon. This is evident when comparing column 4 and 5, which contain the effects for previously non-reviewed listings. The results from the coupon experiment suggest that eliminating sorting by inducing everyone to review would decrease the rate of five star reviews by 6 percentage points, whereas removing strategic motivations only has a 1 percentage point effect. Therefore, sorting is

[16]One worry about the external validity of these results is that not all guests and hosts may have noticed the information about changes to the review system. If knowing about the 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. In Figure A3 we display the long-run trends in ratings as a share of all reviews for a set of exprienced 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 does drop after the public launch, due to some combination of the fact that two-thirds of trips became eligible for the simulatenous reveal system and because of the attention garnered by a blog post and news. However, the long-run ratings trend does not fall substantially after the initial launch, suggesting that attention was not a primary driver of the results.

more important for determining the distribution of star ratings than strategic factors.

Turning to the host related statistics in Table 7, the rate of reviews increases by 7 percentage points, demonstrating that hosts were aware of the experiment and were induced to review. Furthermore, the rate of positive recommendations by hosts increased by 1 percentage point, suggesting that the recommendation is not affected by strategic motivates. However, the text of the submitted reviews does change. The rate of negative sentiment conditional on a non-recommend (calculated using the methodology described in section 2) increases from 71% to 74%. This suggests that the experiment had the intended effect of allowing people to be more honest in their public feedback. Table 8 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 decreases by 3 percentage points. 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 the setup in subsection 2.2, where review information is used to predict rebooking and customer service call rates. We then add an interaction between the rating and the treatment. The results are presented in Table 9. 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.

5.1 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. In the treatment group, the first review text should have no effect on the second review, conditional on the host's recommendation. Our specification to test this is:

$$y_{gl} = \alpha_0 t_l + \alpha_1 FNR_{gl} + \alpha_2 FNS_{gl} + \alpha_3 t_l * FNR_{gl} + \alpha_4 t_l * FNS_{gl} + \beta' X_{gl} + \epsilon_{gl} \qquad (4)$$

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.[17]

Table 10 displays estimates of Equation 4 for cases when the guest reviews second. Columns (1) - (3) show the estimates for guest non-recommendations, low ratings, and negative sentiment 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 do 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. This positive effect reflect 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.

5.2 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:

$$y_{gl} = \alpha_0 t_l + \alpha_1 NE_{gl} + \alpha_2 NE_{gl} * t_l + \epsilon_{gl} \tag{5}$$

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 by the reviewer. We expect α_0 and α_2 to be positive because first reviews should be more honest in the treatment.

Table 11 displays estimates of Equation 5 for first reviews by hosts. Column (1) shows that hosts are 2.7 percentage points 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 eight percentage points more likely to review first when they did contact customer support. Column (3)

[17]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.

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.[18] Conditional on leaving a negative review, hosts leave an additional .2 negative words in the review. These results demonstrate that hosts are aware of strategic considerations and omit negative feedback from public reviews even if they have a negative experience. Appendix D discusses the analogous results for guests reviewing first.

6 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 negative text 14% of the time and leave four or five star ratings 20% 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 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 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. Guests often mention that it feels awkward to leave a negative review after interacting with a host. For example, one guest said: "I liked the host so felt bad telling him more of the issues." Second, guests frequently mention that they don't want the host to feel bad. One respondent said: "I often don't tell the host about bad experiences because I just don't want to hurt their feelings". Third, guests don't want to hurt the host's reputation. A typical response is: "My hosts were all lovely people and I know they will do their best to fix the problems, so I didn't want to ruin their reputations." Lastly, guests sometimes doubt their own judgment of the experience. For example, one guest claimed that "I think my expectations were too high".

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. Our first proxy for the degree of social interaction is whether the trip was to a private room within a home or to an entire property. Stays in a private room are more likely to result in social interaction with the host because of shared space. Our second proxy for social interaction is whether the host is a multi-listing host (defined as a host with more than 3 listings). Multi-listing hosts are less likely to interact with guests because they are busy managing other properties

[18] 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.

and because they typically do not reside in the properties they manage.

Figure 7 plots the distribution of guest ratings conditional on not recommending the host as a function of property type. Guests staying with casual hosts are over 5% 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.

Of course, 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, sometimes a particular property is rented out fully while other times just a room in that property is rented out. 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. Similarly, stays with 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. This variation exists because hosts sometimes start as casual hosts but then expand their operations over time.

Consider the following regression specification:

$$y_{glt} = \alpha_0 PR_l + \alpha_1 MLH_l + \alpha_2 R_{glt} + \alpha_3 MLH_l * NR_{glt} + \alpha_4 R_{glt} * NR_{glt} + \beta' X_{gl} + \gamma_g + \epsilon_{gl} \quad (6)$$

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 a vector of rating indicators, 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 12 displays the results of regressions predicting whether a review rating had more than 3 stars. Columns (1) contains a specification with a variety of controls 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 4.5% 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 3.2 percentage points less likely to receive high rating if the guest does not recommend.

Table 13 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 4.6 percentage point decrease in the probability of high ratings for entire properties relative to private rooms conditional on a non-recommendation. We also conducted the same exercise when the outcome variable was negative sentiment in review text and the results were similar. This evidence confirms that guest's willingness to be honest in reviews is a function of the degree of social interaction they had with the host. Furthermore, 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.

7 Measuring the Size of Bias

Our analysis has shown that submitted reviews on Airbnb exhibit bias from sorting, strategic reciprocity, and socially induced reciprocity. In this section, we describe a methodology for using experimental estimates to measure bias and quantify the relative importance of the mechanisms documented in this paper.

Our first measure of bias, B_{avg}, is the difference between average experience and the reported experience. This metric represents how far the reputation system is on average in representing the experience of users. Our second measure of bias, B_{neg}, is the share of those with negative experiences who reported negatively. This rate quantifies how many bad guests or hosts are "caught". To the extent that a bad agent imposes a negative externality on other agents (Nosko and Tadelis (2015)), the platform may especially care about catching these bad agents in the review system. Furthermore, the welfare losses from imperfect reputation systems in section 3 are a direct function of B_{neg}.

7.1 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. Then the share of submitted ratings that are positive is:

$$\bar{s} = \frac{gPr(r|g)Pr(s_g|g,r) + (1-g)Pr(r|n)Pr(s_g|n,r)}{Pr(r)} \tag{7}$$

where r is an indicator for whether a review was submitted. The difference between the average actual experience and the average submitted review is:

$$B_{avg} = (1-g)\frac{Pr(r|n)Pr(s_g|n,r)}{Pr(r)} - g(1 - \frac{Pr(r|g)Pr(s_g|g,r)}{Pr(r)}) \tag{8}$$

The first term is the share of reviewers with bad experiences who report positively and the second term is the share of all guests with positive experiences who report negatively. Note, these two forms of bias push the average in opposite directions. So looking at average ratings understates the amount of misreporting. However, given that retaliation happens infrequently, this second term should not affect bias much.

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

$$B_{neg} = 1 - \frac{N_{n|n}}{N_{all}(1-g)} \tag{9}$$

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 honestly recommend when they leave a review (because the recommendation is anonymous). To calibrate the empirical analogue to g, we need to make assumptions about the degree of selection into reviewing. Because 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 eliminate some of the effect of selection into reviewing. However, because the coupon experiment was only conducted for listings with 0 reviews, we must extrapolate to the sample of all reviews. To do so, we assume that the relative bias due to sorting for listings with 0 reviews is the same as the bias due to sorting for the overall sample. We then reweigh the baseline rate of recommendation for listings with 0 reviews by the relative rates of recommendations in the overall sample.

$$\hat{g} = s_{0,ir,sr} \frac{s_{all,sr}}{s_{0,c,sr}} \tag{10}$$

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.

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 in the coupon experiment equal to the rate of positive experiences for guests eligible for the coupon experiment who reviewed in the treatment group of the coupon 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 coupon 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. Lastly, we need to calibrate the review probabilities and mis-reporting rates conditional on leaving a review. We describe how to do so in the next section.

7.2 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 biases (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 simple transformations of the moments in the data. $Pr(\widehat{s_g|}n, r)$ is equal to the empirical rate of positive text without a recommendation. $\hat{g} = 3.68\%$ is our best estimate of the true rate of non-recommended experiences in the data and $Pr(\widehat{r|}n) = \frac{Pr(\widehat{n|}r) * \widehat{P(r)}}{(1-\hat{g})}$. Scenario 3 represent the bias if there was no socially induced reciprocity in the reviewing process. To calculate the review rates in this scenario, we set $Pr(\widehat{s_g|}n, r)$ equal to the adjusted rate of positive text for stays with multi-listing hosts in entire properties. Scenario 4 removes sorting bias from reviews. This corresponds to the rate of non-recommendation if the share of reviewers with non-recommended experiences was equal to the share of guests with non-recommended experiences. The no-sorting calculation keeps the overall review rate equal to the review rate in the simultaneous reveal treatment. Lastly, scenario 5 computes the two measures of bias if everyone submits reviews.

Table 14 displays both measures of bias in each of the five scenarios.[19] We first turn to the case when all biases are present (row 1). In this scenario, positive reviews occur 1.84% more of the time than positive experiences. Furthermore, 70% of non-recommended experiences are not reported in text. Rows 2 and 3 display the effects of removing strategic and social reciprocation. Removing these sources of bias reduces the average bias by .34% and the share of negative reviews missing by 7%. Therefore represent 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 .94 percentage points in average bias and 21 percentage points in the share of negative experiences missing. Sorting is a more important source of bias than strategic and socially induced reciprocity using both measures. Lastly, 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 26 percentage points 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 15% 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. In total, there is a less than 2% difference between the imputed rate of negative experiences and the average rate of negative reviews. Bias due to both strategic and social reciprocity exists comprise a relatively small share of the total. Sorting into reviewing represents the biggest source of bias on the platform. When guests do have negative experiences, they do not report them in review text 66% of the time and in the star ratings 59% of the time. The main causes of this non-reporting is that approximately 30% of guests do not submit a review.

8 Discussion

User-generated reviews are an important component of a well-functioning online marketplace. However, because informative reviews are public goods, the review system may be biased. We use experiments and proprietary data from Airbnb to show that public reviews are informative and typically correspond with private and anonymous ratings. Nonetheless, bias in reviews still occurs. Our analysis documents review bias due to sorting, strategic reciprocity, and socially induced reciprocity. These biases leads to cases when negative experiences are not reported in review text on the website. If the three biases were eliminated, then an additional 28% of negative experiences would be documented in reviews on the website.

Based on our results, the most important bias to tackle in review systems is sorting into reviewing. There are several potential interventions that might reduce sorting bias in

[19] See Table AVII for a measure of bias using five star ratings conditional on a non-recommendation.

review systems. 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 on Uber), strategically offering coupons for reviews, or making the review easier to submit. Second, online marketplaces can display ratings that adjust for bias in the review system. 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. Lastly, 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.

Our theoretical model suggests that a key outcome variable determining the costs of imperfect reputation systems is the share of high type sellers entering the platform. This paper shows that the high share of Airbnb users reporting positive transactions is not explained by bias in the review system. This begs the question of what eliminates low quality sellers from the platform.

There are at least three possible mechanisms that could eliminate low quality listings from the site. 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. In future work, we plan to study the importance of these mechanisms for equilibrium outcomes such as transaction volume, prices, and welfare.

References

Andreoni, James, and Justin M. Rao. 2011. "The power of asking: How communication affects selfishness, empathy, and altruism." *Journal of Public Economics*, 95(7-8): 513–520.

Archak, Nikolay, Anindya Ghose, and Panagiotis G. Ipeirotis. 2011. "Deriving the Pricing Power of Product Features by Mining Consumer Reviews." *Management Science*, 57(8): 1485–1509.

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

Benabou, Roland, and Jean Tirole. 2006. "Incentives and Prosocial Behavior." *American Economic Review*, 96(5): 1652–1678.

Bohnet, Iris, and Bruno S Frey. 1999. "The sound of silence in prisoner's dilemma and dictator games." *Journal of Economic Behavior & Organization*, 38(1): 43–57.

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

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

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

Cullen, Zoe, and Chiara Farronato. 2015. "Outsourcing Tasks Online: Matching Supply and Demand on Peer-to-Peer Internet Platforms."

Dai, Weijia, Ginger Jin, Jungmin Lee, and Michael Luca. 2012. "Optimal Aggregation of Consumer Ratings: An Application to Yelp.com."

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.

DellaVigna, Stefano, John A. List, and Ulrike Malmendier. 2012. "Testing for Altruism and Social Pressure in Charitable Giving." *The Quarterly Journal of Economics*, 127(1): 1–56.

Fradkin, Andrey. 2014. "Search Frictions and the Design of Online Marketplaces."

Friedman, Thomas. 2014. "And Now for a Bit of Good News . . ." *The New York Times*.

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

Hoffman, Elizabeth, Kevin McCabe, Keith Shachat, and Vernon Smith. 1994. "Preferences, Property Rights, and Anonymity in Bargaining Games." *Games and Economic Behavior*, 7(3): 346–380.

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

Lazear, Edward P., Ulrike Malmendier, and Roberto A. Weber. 2012. "Sorting in Experiments with Application to Social Preferences." *American Economic Journal: Applied Economics*, 4(1): 136–163.

Li, Lingfang (Ivy), and Erte Xiao. 2014. "Money Talks: Rebate Mechanisms in Reputation System Design." *Management Science*, 60(8): 2054–2072.

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

Malmendier, Ulrike, and Klaus Schmidt. 2012. "You Owe Me." National Bureau of Economic Research Working Paper 18543.

Malmendier, Ulrike, Vera L. te Velde, and Roberto A. Weber. 2014. "Rethinking Reciprocity." *Annual Review of Economics*, 6(1): 849–874.

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

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

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

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

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

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

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.

9 Figures

Figure 1: Review flow on the website

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

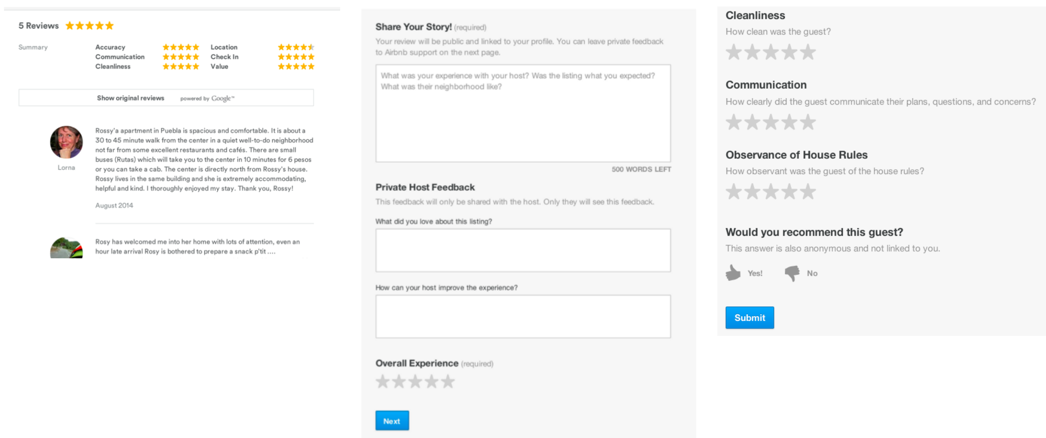


Figure 2: Distribution of Guest Overall Ratings of Listings

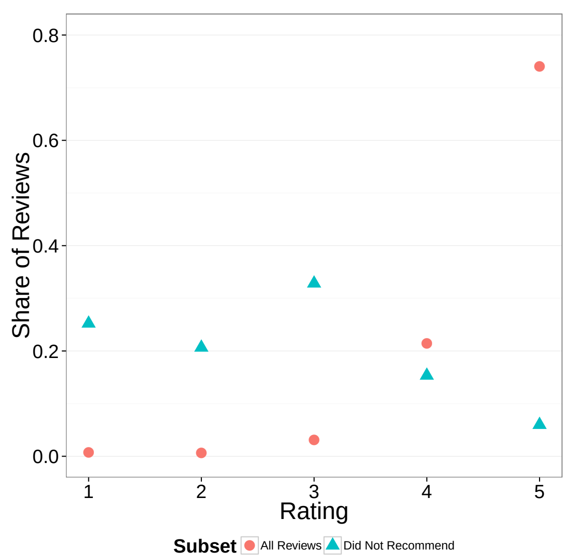
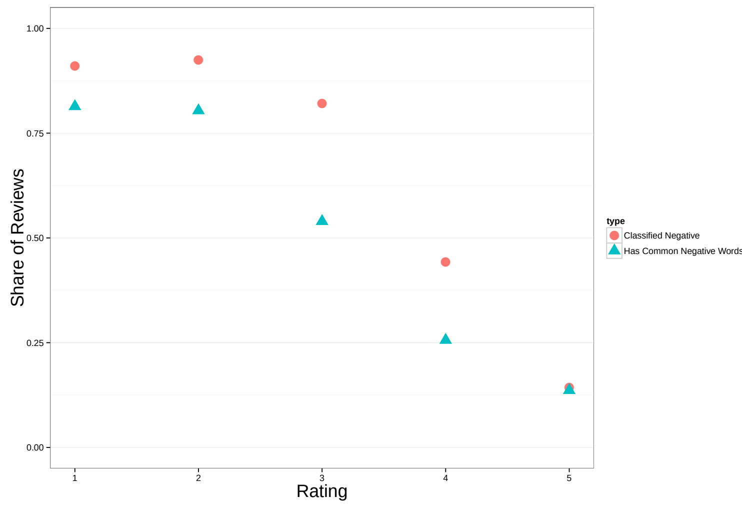


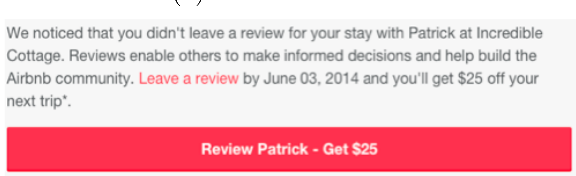
Figure 3: Prevalence of Negative Text Conditional on Rating



"Classified Negative" refers to the classification by the regularized logistic regression based on the textual features of a review. "Has Common Negative Words" is a binary indicator for whether the review contains a word or phrase that occurs in at least 1% of non-recommended reviews and occurs at least 3 times as frequently in guest reviews with non-recommendations as in guest reviews with five star ratings.

Figure 4: Incentivized Review Experiment Emails

(a) Treatment Email



(b) Control Email

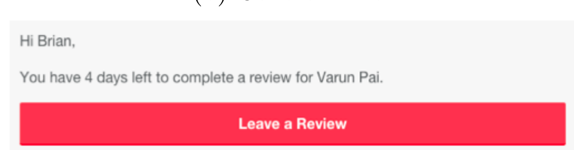


Figure 5: Distribution of Ratings - Experiments

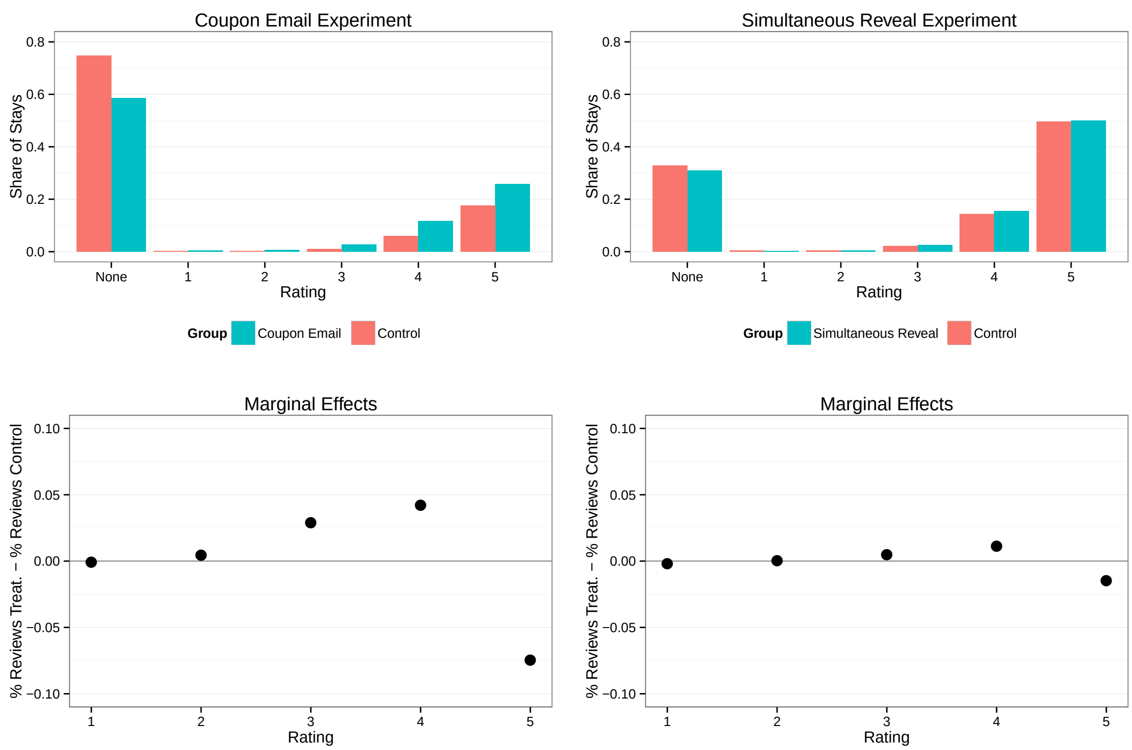
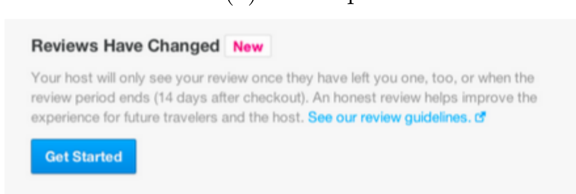


Figure 6: Simultaneous Reveal Notification

(a) Desktop



(b) Mobile

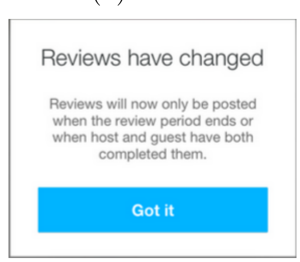
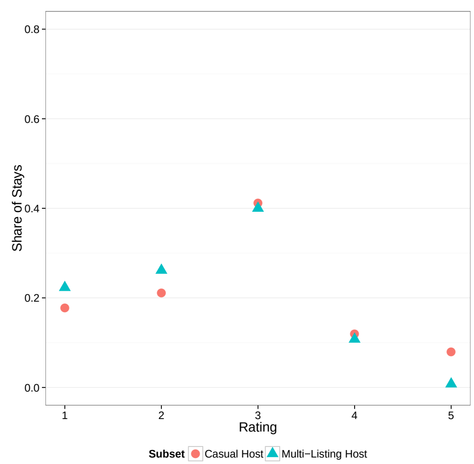


Figure 7: Ratings When Guest Does Not Recommend - Simultaneous Reveal



10 Tables

Table 1: Summary Statistics

Reviewer	Reviews	Five Star	Recommends	Text Classified Positive	Negative Sentiment | Non-Recommend	Time to Review (Days)
Guest	0.671	0.741	0.975	0.779	0.861	4.284
Host	0.715	-	0.989	0.966	0.744	3.667

Table 2: The Informativeness of Reviews: Re-booking Rates

	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

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.

Table 3: Summary Statistics: Incentivized Review Experiment

	Guest		Host	
	Control	Treatment	Control	Treatment
Reviews	0.257	0.426	0.632	0.626
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)	5.729	5.210	2.580	2.618
Negative Sentiment Given Not-Recommend	0.757	0.688	0.948	0.939
Text Classified Positive	0.882	0.930	0.806	0.838
Median Private Feedback Length (Characters)	131	126	96	95
First Reviewer	0.072	0.168	0.570	0.599
Time to Review (Days)	18.420	13.709	5.715	5.864
Time Between Reviews (Hours)	292.393	215.487	-	-
Num. Obs.	15470	15759	15759	15470

Table 4: Magnitudes of Experimental Treatment Effects

Experiment:	**Coupon**	**Coupon**	**Sim. Reveal**	**Sim. Reveal**	**Coupon**
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

Columns (1), (3), and (4) display treatment effects in a linear probability model where the dependent variable is listed in the first column. 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 (5) displays predicted effects on reviews if everyone reviewed. Controls for trip and reviewer characteristics include: number of guests, nights, checkout date, guest origin, listing country, and guest experience. The regressions predicting five star reviews, recommendations, and sentiment are all conditional on a review being submitted. "Negative sentiment" is an indicator variable for whether the review text was classified as negative. *p<0.10, ** p<0.05, *** p<0.01 (Estimates in Column (2) do not have associated standard errors.)

Table 5: Effect of Coupon Treatment on Five Star Ratings

	(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

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

Table 6: Selection into Reviewing: Guest Rebooking Rates

	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

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

Table 7: Summary Statistics: Simultaneous Reveal Experiment

	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	0.765	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	0.779	0.764	0.966	0.964
Negative Sentiment Given Not-Recommend	0.861	0.866	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 8: Cross Tabulation of Outcomes: Simultaneous Reveal Experiment

(a) Control

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

(b) Treatment

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

Table 9: The Effect of Simultaneous Reveal on Rating Informativeness

	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

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

Table 10: Retaliation and Induced Reciprocity - Guest

	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 11: Fear of Retaliation - Host

	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 12: Socially Induced Reciprocity - Star Rating

	Rating > 3		
	(1)	(2)	(3)
Entire Property	−0.011***	−0.013***	
	(0.001)	(0.002)	
Listing Reviews	−0.0001***	−0.0001***	−0.00002
	(0.00000)	(0.00001)	(0.00002)
Checkout Date	−0.000***	−0.000***	−0.000***
	(0.000)	(0.000)	(0.000)
Nights	0.0003***	0.0003***	0.0001***
	(0.00002)	(0.00003)	(0.00005)
Guests	−0.003***	−0.001***	−0.002***
	(0.0001)	(0.0002)	(0.0003)
Customer Support	−0.026***	−0.025***	−0.020***
	(0.001)	(0.001)	(0.001)
Total Bookings by Guest	0.0004***	−0.0002***	−0.0002**
	(0.00003)	(0.0001)	(0.0001)
Price	0.0001***	0.0001***	−0.00003***
	(0.00000)	(0.00000)	(0.00001)
Effective Positive Percentage	0.055***	0.055***	−0.009***
	(0.001)	(0.001)	(0.001)
No Trips	0.003	0.007	0.028
	(0.008)	(0.010)	(0.020)
Person Capacity	−0.001***	−0.001***	−0.0001
	(0.0001)	(0.0001)	(0.0004)
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

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. *p<0.10, ** p<0.05, *** p<0.01

Table 13: Socially Induced Reciprocity - Address Fixed Effects

	Rating > 3		
	(1)	(2)	(3)
Entire Property	0.0005	-0.046^{***}	-0.046^{***}
	(0.002)	(0.005)	(0.007)
Listing Reviews	0.0001^{**}	0.00005	0.00001
	(0.00003)	(0.00003)	(0.0001)
Checkout Date	-0.000^{***}	-0.000^{***}	-0.000^{**}
	(0.000)	(0.000)	(0.000)
Nights	0.0001	0.0001	0.0001
	(0.0001)	(0.0001)	(0.0001)
Guests	0.001	-0.0005	-0.0003
	(0.0005)	(0.0004)	(0.001)
Customer Support	-0.075^{***}	-0.023^{***}	-0.022^{***}
	(0.002)	(0.002)	(0.003)
Log(Guest Bookings)	-0.002^{***}	0.002^{***}	-0.001
	(0.001)	(0.0005)	(0.001)
Log(Price Per Night)	-0.019^{***}	-0.008^{***}	-0.008^{***}
	(0.002)	(0.002)	(0.002)
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

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). *p<0.10, ** p<0.05, *** p<0.01

Table 14: Size of Bias
(Guest does not recommend listing but omits negative text.)

	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

The above table displays three 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 Classifying Text Sentiment

In this section we describe the procedure used to classify review 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".[20] 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%. The top five positive features are "amazing apartment", "fantastic apartment", "of shower", "excellent stay", and "exceeded". The top five negative features were "rude", "ruined", "not clean", "not very clean", and "christmas".

B Predictors of Review Rates

Table AI 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 market city of listing 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.

C Experimental Validity

This section documents that both experimental designs in this paper are valid. Table AIII 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) and 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 for the simultaneous reveal experiment,. However, there is .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. We do not view this difference as a problem because we find balance on all observables and the overall difference in observations is tiny.

D Additional Results on Strategic Reciprocity

In this appendix 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 (< 4 stars) by a guest in the control is associated with a 27 percentage points lower recommendation rate and a 32 percentage points 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

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

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.

Table AV displays the results for fear of retaliation when guests review first. Column (1) shows that there is no difference in whether guests recommend in the treatment and control. Columns (2) and (3) display the effects of the treatment on the likelihood that guests leave a low rating and negative sentiment in their reviews of hosts. There is an overall increase in lower rated reviews by .4 percentage points and an increase in negative sentiment of 1.1 percentage points. Furthermore, column (4) shows that the effect of the treatment does not vary by the quality of the trip, as measured by recommendation rates and ratings. We interpret this small effect as follows. Although guests may fear retaliation, they may have other reasons to omit negative feedback. For example, guests may feel awkward about leaving negative review text or they may not want to hurt the reputation of the host.

One piece of evidence supporting this theory comes from the effect of the treatment on private feedback. 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 AVI displays the effect of the treatment on whether a guest leaves a suggestion. Column (1) shows that the overall effect of the treatment is 6.3 percentage points, suggesting that guests are indeed motivated by fear of retaliation. 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 especially large for guests that recommend the host. Therefore, the treatment allows guests who have good, but not great, experiences to offer suggestions to the host without a fear of retaliation. In the next section we further explore behavioral reasons for reviewing behavior.

E Additional Tables

Table AI: Determinants of Guest Reviews

	Reviewed	
Five Star Rate	0.106***	0.106***
	(0.008)	(0.008)
Past Booker	0.058***	0.058***
	(0.004)	(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	

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.

Table AII: The Informativeness of Reviews: Customer Support

	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

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.

Table AIII: Experimental Validity Check

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	
Multi-Listing Host	Simultaneous Reveal	0.001	0.082	0.081	0.369	
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	
Multi-Listing Host	Incentivized Review	0.002	0.169	0.167	0.678	
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	

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

Table AIV: Retaliation and Induced Reciprocity - Host

	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

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

Table AV: Fear of Retaliation - Guest

	< 5 Rating (First)		Neg. Sentiment (First)	
	(1)	(2)	(3)	(4)
Treatment	0.001	0.003	−0.001	−0.001
	(0.004)	(0.004)	(0.005)	(0.005)
Guest Customer Support		0.115***	0.107***	0.107***
		(0.033)	(0.037)	(0.037)
Non-Recommend		0.673***	0.630***	0.630***
		(0.009)	(0.015)	(0.015)
Treat. * Customer Support		−0.016	−0.016	−0.016
		(0.047)	(0.050)	(0.050)
Treat. * Non-Recommend		−0.030**	−0.018	−0.018
		(0.014)	(0.021)	(0.021)
Guest, Trip, and Listing Char.	Yes	Yes	Yes	Yes
Observations	41,880	38,023	29,546	29,546

The regressions in columns (2) - (4) are estimated only for cases when the guest reviews first. "Treatment" refers to the simultaneous reveal experiment. 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. *p<0.10, ** p<0.05, *** p<0.01

Table AVI: Determinants of Private Feedback Increase

	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

"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

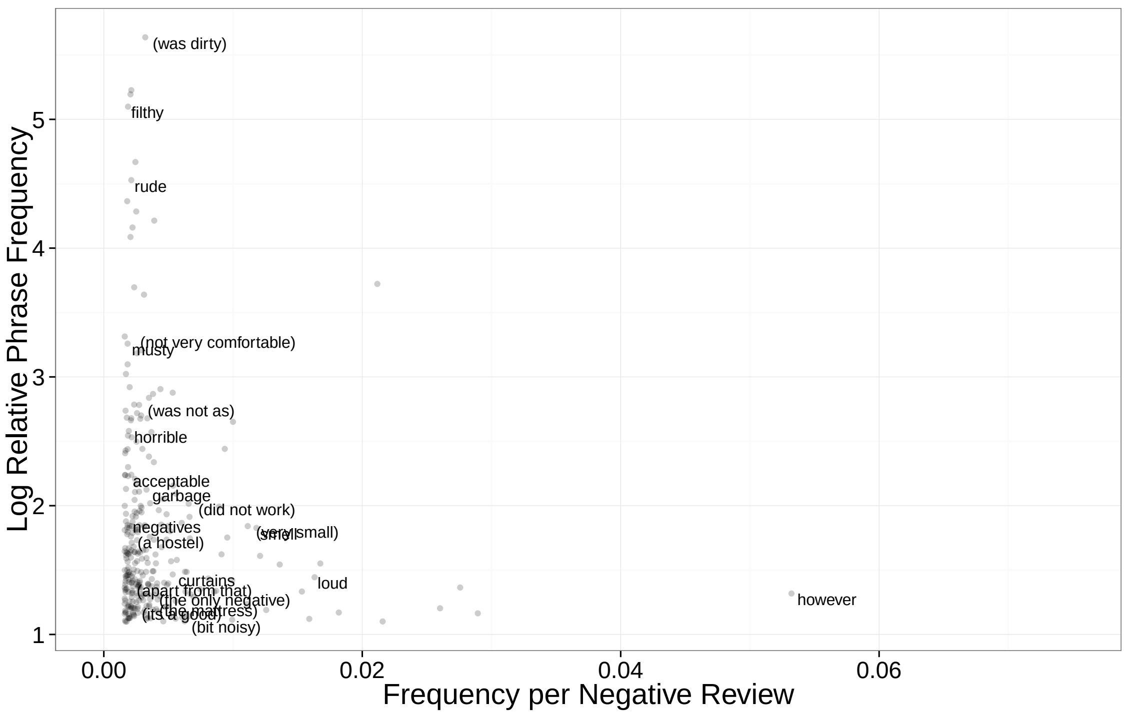
Table AVII: Size of Bias
(Guest does not recommend listing but submits five star rating.)

	Measure of Bias:	
	B_{avg}	B_{neg}
Counterfactual:	Average	% 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

The above table displays three 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_{mis} is the share of all reviews that are mis-reported and B_{neg} is share of all stays where a negative experience was not reported.

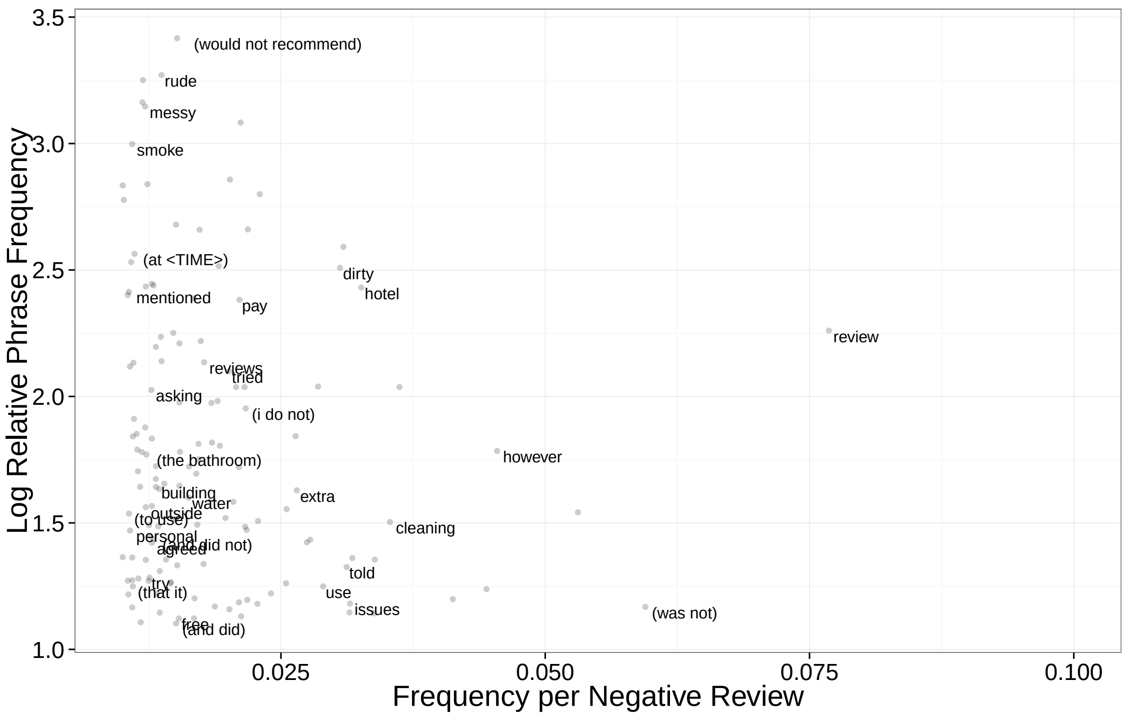
F Additional Figures

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



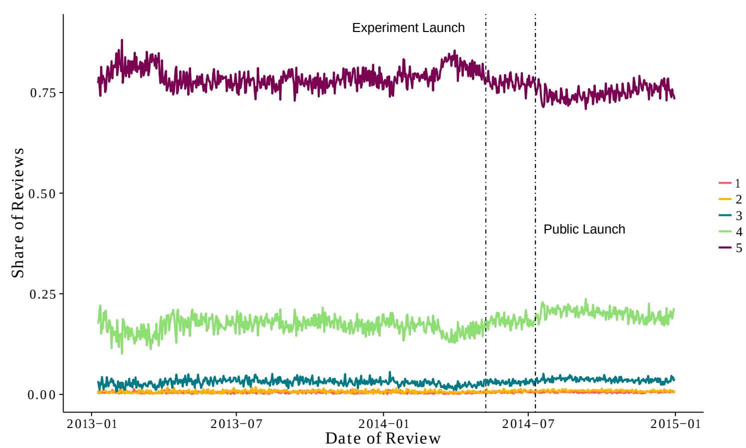
“Relative phrase frequency” refers to the ratio with which the phrase occurs in reviews with a rating of less than five stars.

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



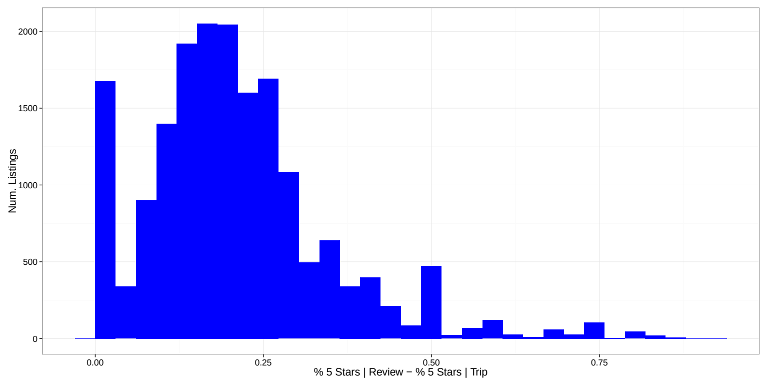
"Relative phrase frequency" refers to the ratio with which the phrase occurs in reviews with a non-recommendation.

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 line demarcates the start of the experiments while the second line demarcates the public launch of the simultaneous reveal system, which placed an additional two-thirds of hosts into the simultaneous reveal treatment.

Figure A4: Histogram of Difference in Ratings per Listing



The sample used for this figure is composed of highly rated listings (> 4.75 average overall star rating) with at least 3 reviews. This sample is chosen because Airbnb only displays star ratings after 3 reviews are submitted and rounds the star rating the nearest .five stars. Therefore, the listings in this sample seem the same to guests on the overall star rating dimension.