Bias and Reciprocity in Online Reviews: Evidence From Field Experiments on Airbnb

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

[1]The National Bureau of Economic Research and Airbnb, Inc.
[2]Airbnb, Inc.

April 28, 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 while reviews on Airbnb are generally informative and 97% of guests privately report having positive experiences, bias still exists. Using two field experiments, we show that non-reviewers have worse experiences than reviewers and that strategic reviewing behavior occurs on the site, although the aggregate effect of the strategic behavior is relatively small. Furthermore, we find that some reviewers who privately report negative experiences, nonetheless submit positive public reviews. We attribute much of this behavior to *socially induced reciprocity*, which occurs when buyers and sellers form a social relationship and consequently omit negative information from reviews. Lastly, we use a quantitative exercise to show that when negative experiences do occur, they are not captured in review text 63% of the time.

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

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, 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

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 in their reviews. 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 (2014) or to leave a platform. Furthermore, a less biased reputation signal can reduce moral hazard by incentivizing good behavior (Hui et al. (2014), Saeedi et al. (2015)). These factors make understanding reviews and other forms of feedback a first order concern for marketplaces that design review systems and organizations that rely on feedback to make decisions.

In this paper we 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. The setting of this paper is Airbnb, a large online marketplace for accommodations. Guests and hosts on Airbnb review each other using publicly displayed text, star ratings, and anonymous recommendations, which are never displayed on the site. 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]

We find that reviews on Airbnb are informative — positive public reviews are correlated with anonymous recommendations and the absence of customer support calls. However, as in other online marketplaces, reviews on Airbnb are predominantly positive.[3] Over 70% of reviewers leave a five star rating and over 97% of reviewers recommend their counterparty. We use field experiments as well as non-experimental results to show that, for both guest and host reviews, bias increases the rates of five star ratings and positive text reviews. Furthermore, although negative experiences are rare, they are not reported in review text over 50% of the time.

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 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

[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, Thomas Friedman 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.

more honest and increase review rates. In our model, the best review system occurs if each agent submits an honest report after a transaction.

We use our first field experiment, described in more detail in section 4, to study selection into reviewing. The experimental treatment offers a $25 coupon in exchange for a review. The treatment increases review rates by 6.4 percentage points and decreases the share of those reviews that are five stars by 2.1 percentage points. Furthermore, the coupon does not cause guests to change their reviewing style as measured by prior reviewing behavior, suggesting that the effect of the experiment is due to pure selection.

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. Over 20% of guests who anonymously answered that they would not recommend their host nonetheless submitted a public review with a four or five star rating and positive review text.[5] 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. Furthermore, Bolton et al. (2012) provides laboratory evidence that changing to a system in which reviews are hidden until both parties submit a review reduces retaliation and makes markets more efficient.

We conduct the first experimental test of the simultaneous reveal mechanism in an online marketplace and use it to test whether misrepresentation is due to the fact that reviewers are strategic. The treatment increases review rates by guests while decreasing the share of five star reviews by 1.6 percentage points. On the host side, the treatment increases review rates by 7 percentage points but does not affect recommendation rates. We document that strategic motives did affect reviewing behavior in the control group but that those effects had minimal effects on the ratings distribution. Our results are notable for two reasons. First, the small size of the change in the review distribution demonstrates that strategic motives are not a primary determinant of the ratings distribution. Second, in contrast to the prediction in Bolton et al. (2012), the simultaneous reveal mechanism increases review rates.

Mismatch between public and private reviews occurs even in the simultaneous reveal treatment group. In section 6 we use non-experimental evidence to study several explanations for 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

[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.

[5] See Horton (2014) for a similar result regarding Odesk

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.[6]

Lastly, we conduct a quantitative exercise to measure the magnitude of bias in online reviews. Bias occurs when a negative experience does not result in a negative public review. We show that bias decreases the rate of reviews with negative text and a non-recommendation by just .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 63% of the time. We find that half of this effect is caused by the biases discussed in this paper and the other half because not every guest submits a review.

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 services and underutilized assets between buyers and semi-professional sellers. There has been relatively little empirical work on this sector.[7]

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

[6] Airbnb surveys have asked guests why they do not submit a bad review. Here are two representative responses: "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!"

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

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.[8] 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.[9]

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 model of reviewing behavior.[10] 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 (2014) 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 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 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. Lastly, even in an unbiased review system, cognitive constraints may prevent agent from using all of the available review information to make decisions.

Lastly, there are several parallels between the social influences on reviewing behavior

[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 of the site at different rates on the two sites.

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

and the social influences on giving in 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 and Descriptive Statistics

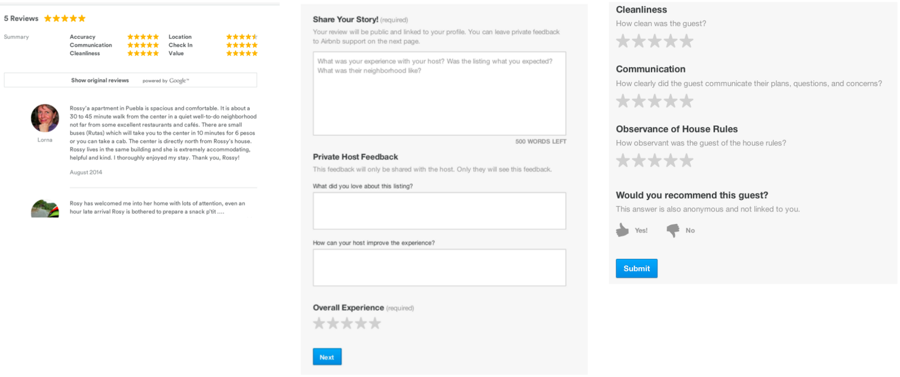
Airbnb describes itself as a trusted community marketplace for people to list, discover, and book unique accommodations around the world. In 2012, Airbnb accommodated over 3 million guests and listed over 180 thousand new listings. Airbnb has created a market for a previously rare transaction: the rental 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 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 3 pages asking public, private, and anonymous questions (shown in Figure 1). Guests are initially asked to leave feedback consisting of publicly shown text, a 1 to five star rating,[11] 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 reviews. Importantly, the second page also contains an anonymous question that asks whether the guest would recommend staying in the listing being reviewed.

[11] 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 browser during most of the sample period.

Figure 1: Review flow on the website



(a) Reviews on Listing Page

5 Reviews ★★★★★

Summary

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

Lorna

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

August 2014

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

(b) Review of Listing (Page 1)

Share Your Story! (required)

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

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

500 WORDS LEFT

Private Host Feedback

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

What did you love about this listing?

How can your host improve the experience?

Overall Experience (required)

☆☆☆☆☆

(c) Review of Guest (Page 2)

Cleanliness

How clean was the guest?

☆☆☆☆☆

Communication

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

☆☆☆☆☆

Observance of House Rules

How observant was the guest of the house rules?

☆☆☆☆☆

Would you recommend this guest?

This answer is also anonymous and not linked to you.

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 59981 trips between May 10, 2014 and June 12, 2014, which are in the control group of the simultaneous reveal experiment.[12] 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 2.7 days to leave a review and guests taking an average of 3.3 days. Hosts review at higher rates and review first more often for two reasons. First, because hosts receive inquiries from other guests, they check the Airbnb website more frequently than guests. Second, because hosts use the platform more frequently than guests and rely on Airbnb to earn money, they have more to gain than guests from inducing a positive guest review.

We first consider guest reviews of hosts. 97% of guests who submit a review for a listing, recommend that listing in a an anonymous question prompt. This suggests that most guests do have a good experience. Figure 2 shows the distribution of star ratings for submitted

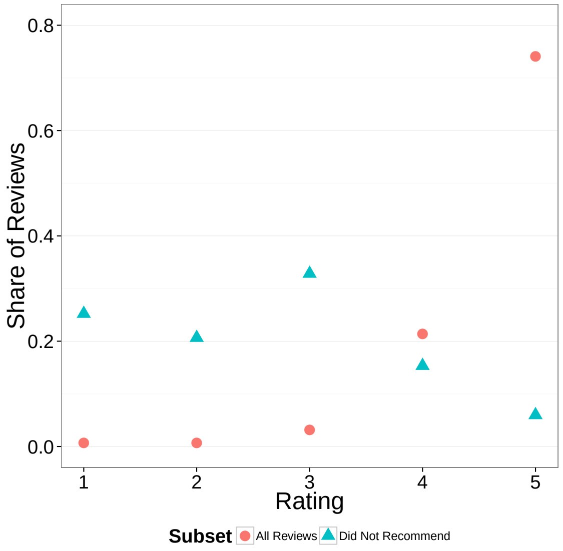
[12] The experiments are randomized at a host level. Only the first trip for each host is included because the experimental treatment can affect the probability of having a subsequent trip. To the extent that better listings are more likely to receive subsequent bookings, these summary statistics understate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

Table 1: Summary Statistics

Reviewer	Reviews	Five Star	Recommends	Overall Rating	First Reviewer	Time to Review (Days)
Guest	0.671	0.741	0.975	4.675	0.350	4.283
Host	0.716	-	0.989	-	0.492	3.667

reviews. Guests submit a five star overall rating 74% of the time and a four star rating 20% of the time.[13]

Figure 2: Distribution of Guest Overall Ratings of Listings



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. Review text is also important because it influences consumer decisions even conditional on numerical ratings (Archak et al. (2011)). Appendix Figures 7 and 8 show phrases that were at least four times as likely to occur in reviews of listings with a lower than five star rating. Phrases that commonly show up in these reviews concern cleanliness, smell, unsuitable furniture, noise, and sentiment.

Figure 2 displays the star ratings conditional on whether a guest recommended the listing. As expected, 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 are sometimes misrepresenting the quality of their experiences in star ratings. One worry is that guests do not understand the review prompt. Although we have no way to test this directly, we are reassured by the fact that fewer than 5% of reviewers recommend

[13] 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.

a listing when they submit a lower than four star rating.[14]

Negative review text should be even less prevalent than low star ratings because text is public, personally linked, and requires the most effort to write. We detect negative sentiment in English language reviews by checking whether the review text contains any of the words or phrases displayed in Figure A1. This procedure results in some classification error because some words labeled as negative such as "however" might be used in a positive review. Alternatively, some infrequent but negative words may not be identified by the procedure. Nonetheless, this strategy is informative about review content. We find that over 80% of 1 and 2 star reviews contain at least one of the negative phrases. However, only 53% of 3 star reviews and 24% of 4 star reviews contain negative sentiment. Therefore, even guests who are willing to leave a lower star rating are often unwilling to submit a negative public review. We examine this phenomenon in greater detail in subsection 5.1.

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

When bad events do occur, host reviews are partially informative. For example, hosts' recommendation rates fall by 11 percentage points for stays in which hosts contact customer support. We use hosts' review texts to determine the characteristics of a bad guest (Appendix A2). Negative reviews contain phrases concerning sentiment, personal communication, money, cleanliness, and damage.

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 equal probabilities of being a high type or low 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 $\underline{u} \leq \frac{2u_l+u_h}{3}$. 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.

[14] Both guests who recommend and don't recommend submit 4 star ratings over 20% of the time. This suggests that a 4 star rating may mean different thing depends on who submitted a review. This source of heterogeneity is an important consideration in designing reputation systems, but we do not consider it further in this paper.

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 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 utility of being positive in a review and is influenced by social and strategic reciprocity and the overall 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.

This is an implication of the fact that not everyone reviews and that not everyone tells the truth about a low quality experience. One argument against the generality 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 .five stars and sometimes even good sellers get bad reviews. Therefore, buyers still face situations where multiple seller types have the same rating.

Observation 2: The platform knows more information than buyers 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' recommendations and customer service calls.

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 $(\alpha_i, \, \beta_i, \, \gamma)$.

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:

$$Welfare = .5u_h + .5u_l(1 - .5r_{ll}) - \underline{\text{u}}(1 - .25r_{ll}) \tag{2}$$

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 $.25(\underline{\text{u}} - u_L)(1 - r_{ll})$.

Therefore, the gain from having a better review system is a function of receiving honest 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}.

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. Therefore, the welfare 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.

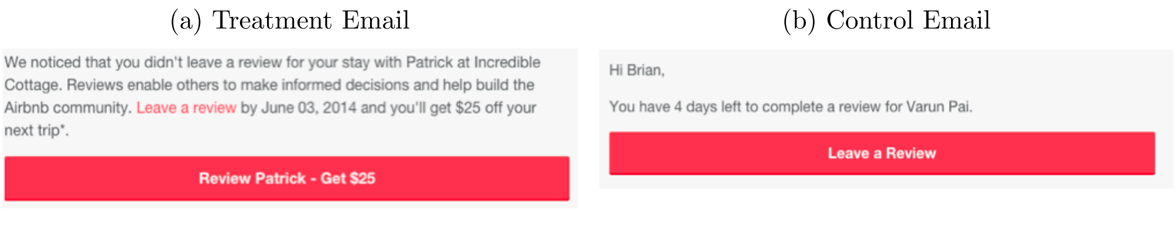
Increasing α_i induces additional buyers to misreport and induces truth-tellers to misreport. The welfare losses from increasing α_i come from inducing truth-tellers to dishonestly report because only r_{ll} matters for welfare. Tying this to our later empirical results, increasing socially induced reciprocity 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 Sorting Into Reviewing and the Incentivized Review Experiment

In this section we describe an experiment and use it to study the magnitude of sorting based on the quality of experience by reviewers. Sorting occurs when the quality of a user's experience is related to the probability of a review. Our experiment induces additional guests to submit reviews. We should that the additional reviews induced by the experiment are worse than the reviews in the control.

The experiment we study offers was ran by Airbnb in 2014 in order to induce reviews for non-reviewed listings. 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 (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 3).

Figure 3: Incentivized Review Experiment Emails



(a) Treatment Email

We noticed that you didn't leave a review for your stay with Patrick at Incredible Cottage. Reviews enable others to make informed decisions and help build the Airbnb community. Leave a review by June 03, 2014 and you'll get $25 off your next trip*.

Review Patrick - Get $25

(b) Control Email

Hi Brian,

You have 4 days left to complete a review for Varun Pai.

Leave a Review

The treatment affected the probability of a review and consequently the probability of additional bookings for a listing. This resulted in more trips 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.[15] Appendix B demonstrates that the randomization for this experiment is valid.

Figure 4: Distribution of Ratings - Experiments

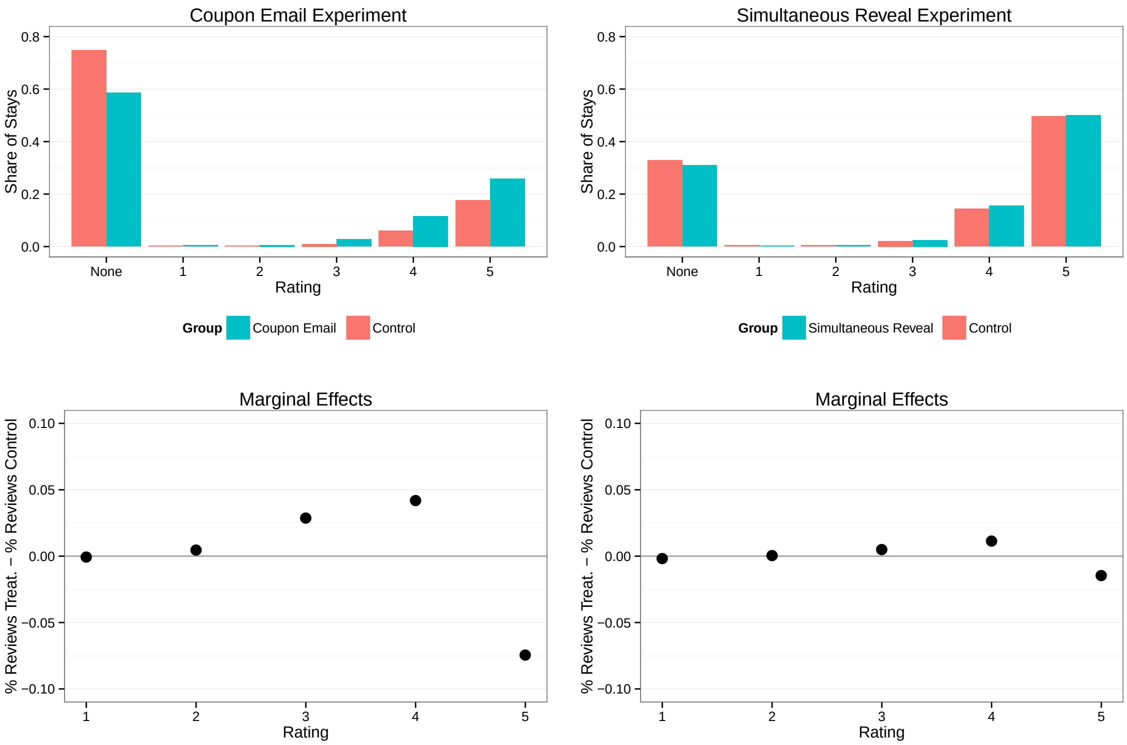


Table 2 displays the review related summary statistics of the treatment and control groups in this experiment. First, note that the 23% review rate in the control group is smaller than the overall review rate (67%). The lower review rate is due to the fact that those guests who do not review within 9 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 11%. The left panel of figure 4 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.

The effect of this experiment on the review ratings might be caused by one of several mechanisms. In Appendix C we show that the effect is not driven by guest characteristics, guest leniency in reviewing, listing characteristics, and fear of retaliation in a host review.

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

[15] We plan to investigate what occurred in subsequent trips in follow-up work.

Table 2: Summary Statistics: Incentivized Review Experiment

	Control		Treatment	
	Guest	Host	Guest	Host
Reviews	0.257	0.627	0.426	0.632
Five Star	0.687	-	0.606	-
Recommends	0.773	0.985	0.736	0.986
High Likelihood to Recommend Airbnb	0.730	-	0.707	-
Overall Rating	4.599	-	4.488	-
All Sub-Ratings Five Star	0.457	0.795	0.389	0.805
Responds to Review	0.021	0.051	0.019	0.040
Private Feedback	0.431	0.273	0.439	0.274
Feedback to Airbnb	0.102	0.089	0.117	0.088
Median Review Length (Characters)	344	126	300	128
Negative Sentiment Given Not-Recommend	0.203	-	0.217	-
Median Private Feedback Length (Characters)	131	95	126	97
First Reviewer	0.072	0.599	0.168	0.570
Time to Review (Days)	18.394	5.861	13.705	5.715
Time Between Reviews (Hours)	291.601	-	215.562	-
Num. Obs.	15430	15430	15719	15719

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.

$$e = \frac{s_{\leq 9} r_{\leq 9} + (s_{ctr} + t_{rev})(r_{ctr} + t_r)}{s_{\leq 9} + s_{ctr} + t_{rev}} - \frac{s_{\leq 9} r_{\leq 9} + s_{ctr} r_{ctr}}{s_{\leq 9} + s_{ctr}} \tag{3}$$

where e is the adjusted treatment effect for all reviews, s refers to the share of trips in each group, t refers to the experimental treatment effect, and r refers to the mean value of a review metric. "≤ 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 3: 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	0.017***	0.007	0.383
Five Star	-0.122***	-0.023	-0.015***	-0.010	-0.058
Recommends	-0.014	-0.004	-0.001	-0.001	-0.012
Neg. Sentiment	0.051	0.009	0.020***	0.029***	0.020

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. 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 contains one of the phrases identified as negative. *p<0.10, ** p<0.05, *** p<0.01 (Estimates in Column (2) do not have associated standard errors.)

Table 3 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). 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 2.1 percentage points and the share of reviews with recommendations by .5 percentage points.

The effects discussed above do not capture the full bias due to sorting because the experiment induced only 6.4% of guests to review, leaving 27% of trips without guest reviews. Our data cannot tell us about the experiences of those non-reviewers. On the one hand, those who do review in the treatment may have even worse experiences than reviewers because they did not bother to take the coupon. Alternatively, those who did not review may have simply been too busy to review or may have cared about a $25 coupon for Airbnb, especially if they do not plan on using Airbnb again. In column (5) of Table 3 we show the imputed selection effect if non-reviewers had the same behavior as reviewers in the treatment group of the experiment. In this case, there would be a 5.8 percentage point lower five star review rate for previously non-reviewed listings on Airbnb. We view this as conservative estimate of total sorting bias because those not induced by the coupon are likely to have had even worse experiences than those who did take up the coupon.

Although we have documented sorting bias, this bias may not matter much for market efficiency if it is constant across listings. In that case, consumers can rationally adjust their expectations that reviews are inflated. However, if sorting differs between listings with similar ratings then even knowledgeable guests may mistakenly book a low quality listing because that listing has especially biased reviews. We demonstrate that there is heterogeneity in bias by comparing the distribution of the difference between the two quality measures: the share of five star reviews out of all reviews (which does not take into account sorting) and the more reliable Effective Positive Percentage (EPP) proposed by Nosko and Tadelis (2014).[16] We use a sample of listings where the average star rating is greater than 4.75, so that the overall star ratings are rounded to 5 for all listings. Although all of these listings have similar average star ratings, their five star rate minus EPP varies greatly, with an interquartile range for the difference of 16% - 27%. (See Figure A3 for a histogram). In Appendix A we also show that EPP at the time of bookings predicts future ratings. Therefore, we conclude that sorting bias varies across listings and therefore affects market efficiency.

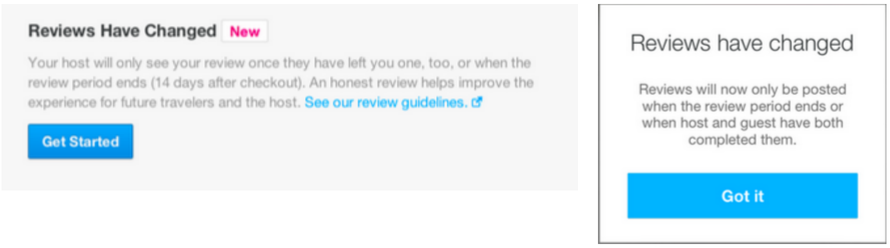
5 The Simultaneous Reveal Experiment

Once concern with Airbnb's two-sided review system is that reviewers may respond to each others reviews and that this would induce strategic reviewing behavior. Our second experiment excludes this possibility by changing the timing with which reviews are publicly revealed on Airbnb. Prior to May 8, 2014, both guests and hosts had 30 days after the checkout date to review each other and any submitted review was immediately posted to the

[16] EPP is measured in this paper by the share of five star reviews out of all trips.

website. This setup allowed for the second reviewer to see the first review and to retaliate or reciprocate in response. 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 5). Another third of hosts were assigned to a control group where reviews were revealed as soon as they were submitted and there were also 14 days to review.

Figure 5: Simultaneous Reveal Notification



(a) Desktop

Reviews Have Changed New

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

Get Started

(b) Mobile

Reviews have changed

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

Got it

Our sample for this experiment consists of the first trip to every listing that was in the experiment. We exclude subsequent trips because the treatment may affect re-booking rates. Appendix B documents the validity of our experimental design. Table 4 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. The experiment had a bigger effect on the rate of private feedback to the host (see Table AVI). The treatment induced a 6.4 percentage points higher rate of guest suggestions to hosts. This increase was present across guest recommendation and star ratings. The increase in suggestions suggests that without the fear of retaliation, guests felt they could speak more freely to the hosts, but that they still did not wish to hurt hosts by leaving negative public review.

Columns (3) and (4) of Table 3 display the experimental treatment effects on guest reviews when controlling for trip and guest characteristics. Column (3) uses the entire experimental sample while column (4) shows estimates from a sample of previously non-reviewed listings. 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) to column (5) of the table. The results from the coupon experiment suggest that eliminating sorting by inducing everyone to review would decrease the rate of five star reviews by 5.8 percentage points, whereas removing strategic motivations only has a 1 percentage point effect. Therefore, sorting is more important for determining the distribution of star ratings than strategic factors.

Turning to hosts, 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 recommendations by hosts did not significantly change, 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

Table 4: Summary Statistics: Simultaneous Reveal Experiment

	Control		Treatment	
	Guest	Host	Guest	Host
Reviews	0.671	0.716	0.690	0.787
Five Star	0.741	-	0.726	-
Recommends	0.975	0.989	0.974	0.990
High Likelihood to Recommend Airbnb	0.765	-	0.759	-
Overall Rating	4.675	-	4.660	-
All Sub-Ratings Five Star	0.500	0.855	0.485	0.840
Responds to Review	0.025	0.067	0.066	0.097
Private Feedback	0.496	0.318	0.568	0.317
Feedback to Airbnb	0.105	0.068	0.109	0.072
Median Review Length (Characters)	327	147	333	148
Negative Sentiment Given Not-Recommend	0.677	0.714	0.680	0.742
Median Private Feedback Length (Characters)	131	101	129	88
First Reviewer	0.350	0.492	0.340	0.518
Time to Review (Days)	4.283	3.667	3.897	3.430
Time Between Reviews (Hours)	63.671	-	47.472	-
Num. Obs.	60683	60683	60968	60968

the methodology described in section 2) increases from 71% to 74%. This suggests that the experiment did have the intended effect of allowing people to be more honest in their public feedback.

5.1 Evidence for Retailiation and Reciprocity

In this section, we use experimental variation to quantify the importance of strategic reviewing 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. We first test whether the relationships between the first review and the second review changes due to the experiment. Our specification is:

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

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

Consider the case when a host submits a review and the guest can see it. If guests are induced to leave positive reviews but host's positive 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, if there is retaliation against negative host reviews, we would expect α_2 to be positive and α_4 to be negative because guests in the treatment can no longer see the first review content. Furthermore, we would expect α_2 to approximately equal $-\alpha_4$. Lastly, we expect that the coefficients on whether the host did not recommended the guest, α_1 to be positive and α_3 to be close to 0. α_1 captures the fact that experiences of guests and hosts are correlated, even if there is no retaliation. However, because the recommendation is always anonymous, there should be no effect of the treatment on this relationship.[17]

Table 5 displays estimates of Equation 4 for cases when the guest reviews second. Column (1) shows the estimates when the outcome variable is whether the guest does not recommend the host. The treatment effect in cases when the first review is positive is not statistically different from 0. This demonstrates that guests do not change their non-public feedback in response to positive host reviews. Next, we consider the effect of a host's review having negative sentiment. We define this variable by looking at all cases where the host does not recommend the guest and where one of the phrases in Figure A2 appears in the review text. The coefficient on host negative sentiment is .67 and the interaction with the treatment is -.63. The two effects approximately cancel each other out, demonstrating that guests retaliate against negative text, but only if they see it. Furthermore, the effect on guest recommendations is large compared to the 97% baseline rate of recommendations. Columns (2) and (3) display the same specification for low ratings by guests and for negative sentiment by guests (defined across all reviews regardless of a guest's recommendation). We see the same pattern of retaliation using these outcome variables.[18] Furthermore, the overall treatment effect, α_0, is approximately .03 for both the rating and sentiment regressions. This demonstrates that guests are induced to leave positive public reviews by positive host reviews. However, the effect of induced reciprocity is an order of magnitude smaller than the effect of retaliation on guest reviews. Therefore, we conclude that guests both retaliate and reciprocate host reviews.

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:

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

where y_{gl} is a negative review outcome, t_l is an indicator for whether the listing is in the treatment group and DNR_{gl} is an indicator for whether the reviewer did not anonymously recommended the counter-party. We expect α_0 to be positive because first reviews should

[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.

[18] The size of the retaliatory response is smaller for negative sentiment. This is likely due to a combination of measurement error in the classification of guest reviews and a hesitation of guests to leave public negative reviews.

Table 5: Retaliation and Induced Reciprocity - Guest

	Does Not Recommend	Overall Rating < 5	Negative Sentiment
	(1)	(2)	(3)
Treatment	0.002	0.032***	0.034***
	(0.002)	(0.006)	(0.005)
Host Negative Sentiment	0.675***	0.700***	0.370**
	(0.130)	(0.122)	(0.163)
Host Does Not Recommend	0.131	0.048	0.261**
	(0.094)	(0.109)	(0.132)
Treatment * Host Negative Sentiment	−0.625***	−0.689***	−0.418**
	(0.161)	(0.175)	(0.209)
Treatment * Host Does Not Recommend	−0.013	0.250*	−0.028
	(0.120)	(0.149)	(0.171)
Guest, Trip, and Listing Char.	Yes	Yes	Yes
Observations	18,010	18,010	18,010

be more honest in the treatment. Furthermore, we expect α_2 to be especially strong in cases where the host does not recommend a guest.

Table 6 displays estimates of Equation 5 for first reviews by hosts. Column (1) displays the effect of the treatment on the probability that a host reviews first. 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. Columns (2) and (3) display the main specification, where y_{gl} is an indicator for the presence of negative sentiment in the host's review text. There is only a .2 percentage point increase in the overall rate of negative text in first host reviews. Column (3) shows that this effect is concentrated amongst hosts that do not recommend the guest. The treatment causes hosts to include negative review text an additional 12 percentage points when they do not recommend the guest.

These results demonstrates that hosts are aware of strategic considerations and omit negative feedback from public reviews even if they have a negative experience. Furthermore, the effect we find is concentrated on hosts who do not recommend their guests. This is a large effect, given that hosts omit negative feedback over 50% of the time when they do not recommend. Lastly, those who do recommend have positive enough experiences so that they feel no need to leave negative feedback, even when there is no strategic penalty for doing so. Appendix D discusses the analogous results for guests reviewing first.

Table 6: Fear of Retaliation - Host

	Reviews First	Neg. Sentiment (First)	
	(1)	(2)	(3)
Treatment	0.027*** (0.003)	0.002* (0.001)	−0.0005 (0.001)
Does Not Recommend			0.613*** (0.044)
Treatment * Does Not Recommend			0.117** (0.055)
Guest, Trip, and Listing Char.	Yes	Yes	Yes
Observations	121,459	31,572	31,572

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 33% of the time and leave four or five star ratings 20% of the time. Similarly, hosts do not leave negative text in 29% of cases when they do not recommend the guest. In this section, we link this misreporting in public reviews to the type of relationship 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.

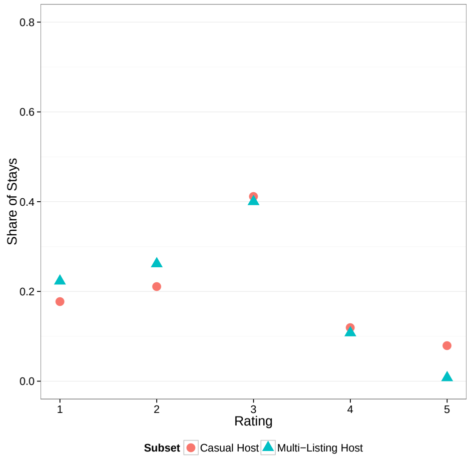
Internal Airbnb surveys of guests suggest that the social aspect of Airbnb affects reviewing behavior. Guests often mention that it feels awkward to leave a negative review after interacting with a host. For example, one guest said: “I liked the host so felt bad telling him more of the issues.” Second, guests frequently mention that they don’t want the host to feel bad. For example, 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 and because they typically do not reside in the properties they manage.

Because trips to different types of listings can differ in quality as well as social interaction, we control for measures of trip quality. Our strategy for identifying the effect of social reciprocity relies on the degree to which there is a mismatch between public and anonymous review ratings. Anonymous ratings should be less influenced by social interactions than public ratings. If socially induced reciprocity occurs, then guests should submit higher public ratings conditional on the anonymous ratings they submit.

Figure 6 graphically shows our identification strategy by plotting 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 publicly.

Figure 6: Ratings When Guest Does Not Recommend - Simultaneous Reveal



Our regression specification to formally test for this effect is:

$$y_{gl} = \alpha_0 PR_l + \alpha_1 PM_l + \alpha_2 R_{gl} + \beta' X_{gl} + \epsilon_{gl} \tag{6}$$

where y_{gl} is a negative review by guest g for listing l, PR_l is an indicator for whether the listing is a private room, PM_l is an indicator for whether the host is a multi-listing host, R_{gl} is a vector of rating indicators, and X_{gl} are guest and trip characteristics. If socially induced reciprocity occurs then we expect α_0 to be negative because guests to private rooms should leave less negative feedback. Furthermore, we expect α_1 to be positive because multi-listing hosts induce less reciprocity in guests.

Table 7 displays the results of the above specification for negative sentiment. Columns (1) - (3) display the coefficients on multi-listing host and private room, with progressively more controls. Column (1) shows that guests staying at a private room are 1.5 percentage points less likely to submit negative review text and guests staying with a multi-listing host are 2.2 percentage points more likely to leave negative text. The effect of ratings on sentiment is in the expected direction, with five star ratings being 50 percentage points less likely to contain negative text. Column (2) adds additional controls for the lowest sub-category rating that a guest submits. The coefficients on room type and multi-listing host barely change in this specification. Lastly, there is a worry that there is measurement error in our classification of review text. In that case, longer reviews may be more likely to be labeled as negative, regardless of their content. To control for this measurement error, column (3) adds controls for a third-degree polynomial in the length of the review. The coefficient on private room remains the same, while the coefficient on multi-listing host increases to 2.8 percentage points.

Table 7: Socially Induced Reciprocity - Negative Sentiment

	Negative Sentiment		
	(1)	(2)	(3)
Entire Property	0.017***	0.018***	0.017***
	(0.005)	(0.005)	(0.005)
Multi-Listing Host	0.020**	0.018*	0.027***
	(0.010)	(0.010)	(0.009)
Guest Does Not Rec.	0.188***	0.150***	0.100***
	(0.024)	(0.025)	(0.022)
Guest, Trip, and Listing Char.	Yes	Yes	Yes
Market FE	Yes	Yes	Yes
Rev. Length Polynomial	No	No	Yes
Rating FE	Yes	Yes	Yes
Subrating FE	No	No	Yes
Observations	29,711	29,711	29,711

We have shown that, holding all else equal, trips with a higher likelihood of social interaction result in more mismatch between public and anonymous ratings. In our preferred specification, column (3), a review of a private room with a casual host is 3.6 percentage points less likely to have negative text than a review of an entire property with a multi-listing host. Furthermore, because trips to entire properties with multi-listing hosts often have social interaction, this estimate is likely to be an underestimate of the true effect of socially induced reciprocity.

There are several concerns with the above identification strategies. First, there may be many differences between listings of different types that may result in differences in reviewing behavior. We address this concern by comparing the behavior of reviewers who stay at the same address but in different room types. This occurs when hosts sometimes rent out a room in the place and other times the entire place. This identification strategy holds the host and most listing characteristics constant. Similarly, we compare stays to a given listing when a host was casual versus when a host was a multi-listing host. In both cases, our estimates do not substantially differ from the baseline effects of mismatch. Second, the results may be driven by the fact that different types of guests stay with different listings and hosts. We address this concern by adding guest fixed effects to the regressions. Appendix E discusses the results from alternate identification strategies.

7 How Large is the Bias?

The goal of the exercise in this section is to quantify the degree of bias caused by each mechanism discussed in this paper. Bias in this exercise refers to cases when a reviewer does not recommend the reviewee but leaves no negative textual feedback.[19] We focus on this measure because it is the clearest case of mis-representation on the website and is prone to the most bias from strategic reciprocity. We ignore cases when guests mis-report positive experiences because retaliations happen fewer than .1% of the time in our sample. Lastly, there are cases when we detect negative text in reviews where the guest recommends the listings. We view these as legitimate positive reviews, with some information that is not positive included. Therefore, we don't count these reviews as mis-reports of a positive experience.

We measure bias for guest reviews of listings in five scenarios, each with progressively less bias. Scenario 1 represent the baseline scenario which represents the control group in the simultaneous reveal experiment. In this case all three biases (sorting, strategic, and social) operate. Scenario 2 removes corresponds to the treatment group of the simultaneous reveal experiment. Note that review rates in this scenario increase where the rate of high ratings decreases. For both scenarios, we calculate measures of bias by making simple transformations of the moments in the data. $\widehat{Pr(s_g|n,r)}$ is equal to the empirical rate of positive text without a recommendation. $\hat{g} = 3.0\%$ is our best estimate of the true rate of negative experiences in the data and $\widehat{Pr(r|n)} = \frac{\widehat{Pr(n|r)} * \widehat{P(r)}}{(1-\hat{g})}$. Scenario 3 further removes social bias in the reviewing process. To do so, we let $\widehat{Pr(s_g|n,r)}$ 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 average experience on the website was equal to the non-recommendation rate for the union of treatment groups in the two experiments. The no-sorting calculation keeps the overall review rate equal to the review rate in the simultaneous reveal treatment. Lastly, scenario 5 computes the measures of bias if everyone reviews.

For each of the 5 scenarios we compute three aggregate bias metrics: B_{avg}, B_{mis}, and B_{neg}.[20] B_{avg} is an estimate of the difference between the share of reviews with positive text in the review system and the share of reviews with positive text that would occur if everyone reviewed and did so honestly. Table 8 displays each measure of bias each scenario. We first turn to the case when all biases are present (row 1). In this scenario, positive reviews occur 1.27% more of the time than positive experiences. Furthermore, .76% of all reviews mis-represent the quality of a guests experience and 65.5% of negative experiences are not reported in text. Removing strategic considerations barley changes these rates of bias. Interestingly, the share of misreported reviews actually increases in the treatment because additional guests with negative experiences were induced to review by the experiment. Therefore, we conclude that strategic motivations have little effect on the rate at which

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

negative experiences are reported in text by guests.

Table 8: Size of Bias
(Guest does not recommend listing but omits negative text.)

	Measure of Bias:		
Counterfactual:	B_{avg} Average	B_{mis} % Misreported	B_{neg} % Negative Missing
Baseline	1.27	0.76	65.45
Simultaneous Reveal	1.22	0.82	63.01
Simultaneous Reveal + No Social Reciprocity	0.76	0.36	53.15
Simultaneous Reveal + No Social Reciprocity + No Sorting	0.41	0.41	40.55
Above + Everyone Reviews	0.41	0.41	13.85

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.

Row 3 shows the bias in the case where social reciprocity is removed as a motivation for reviews. The overall bias and share misreported fall by 37% and 56% respectively. Furthermore, removing social bias while holding review rates constant would decrease the rate of negative experiences missing in the review system by 15%.

In row 4, we remove sorting bias. The effects of removing this bias are similar in magnitude to the effects of removing social reciprocity. Note, B_{avg} and B_{mis} are equivalent in this scenario because we do not consider false negatives in this exercise. Lastly, in Row 5 we report what our measures of bias would be if every guest submitted a review in addition to the scenario in the previos row. B_{avg} and B_{mis} do not change in this scenario because the rate of misreporting does not change. However, B_{neg} falls by an additional 29 percentage points due to the fact that even without sorting, some non-reviewers would have negative experiences. There is a residual 13% of negative experiences that would still go unreported. This misreporting can correspond to two scenarios. First, some of it is due to measurement error. Second, guests may feel socially induced reciprocity even towards property managers and this behavior is not captured in our estimates of socially induced reciprocity.

8 Discussion

There is substantial heterogeneity in ratings and review rates across listings on Airbnb. We have documented that some of that heterogeneity is due to review bias caused by sorting, strategic reciprocity, and socially induced reciprocity. Furthermore, we have shown that although most experiences on Airbnb are positive and baseline bias is low, negative experiences are often not reported in review text on the website. If the three biases were eliminated, then at least 25 percentage points more negative experiences would be documented in review text on the website.

There are at least three ways to alleviate bias in reputation systems. First, marketplaces can change the way in which reviews are prompted and displayed. For example, the simul-

[20] The details for computing these metrics are described in Appendix F

taneous reveal experiment described in this paper eliminated review bias due to retaliation, fear of retaliation, and strategic reciprocity. In fact, after the success of the experiment, the review system was launched to all of Airbnb. Other potential interventions include making reviews mandatory (as on Uber) or changing the review prompt in ways that nudge reviewers to be more honest. 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, review information can be displayed alongside a market level distribution of ratings and the relative rank of the seller being considered. Lastly, as in Nosko and Tadelis (2014), the platform can choose to promote options in search that contain less biased reviews. Furthermore, platforms should consider augmenting review data with other signals of customer experience for the purpose of curation.

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.

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."

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."

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

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. 2014. "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.

A 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.

B Experimental Validity

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

C Robustness Checks for Incentivized Review Experiment

In this section we test for alternative explanations for the effects of the incentivized review experiment. Column (1) of Table AIII displays the baseline treatment effect of the experiment without any control. Column (2) adds in control for guest origin, experience, and trip characteristics. The treatment effects in columns (1) and (2) are approximately equal (-7.5 percentage points), therefore the treatment is not operating by inducing different types of guests to review.

Column (3) shows estimates for a sample of experienced guests and adds controls for the historical judiciousness of a guest when leaving reviews. The guest judiciousness variable measures the extent to which the guest has previously submitted lower ratings. It is equal to the negative guest-specific fixed effect in a regression of ratings on guest and listing fixed effects.[21] As expected, the coefficient on the guest judiciousness term is negative, with pickier guests leaving lower ratings. However, adding this control and limiting the sample to experienced guests does not diminish the effect of the experiment on ratings. Furthermore, the interaction between the treatment and guest judiciousness is not significant. Therefore, the rating behavior of these guests, conditional on submitting a review, does not change due to the presence of a 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 6.9 percentage points. We conclude that the coupon works mainly by inducing those with worse experiences to submit reviews.

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 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

[21] The estimation sample for the fixed effects regressions is the year before the start of the experiment.

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 Results on Socially Induced Reciprocity

In this section we study whether the differences in misreporting by room type and host type are robust to alternate identification strategies. The baseline specification regresses an indicator for whether the guest left a star rating greater than 3 as a function of listing characteristics, market fixed effects, and the private recommendation. The results of this specification are seen in Column (1) of Table AVII. There most relevant stylized fact from this regression is that both entire properties and multi-listing hosts are less likely to recieve a high rating when the guest does not recommend the listing. This difference is 3.1 percentage points for multi-listing hosts and 1.4 percentage points for entire properties. One worry with this specification is that different guests may stay at different types of listings. Column (2) adds guest fixed effects and the parameters of interest do not change.

Lastly, it could be that there are other differences between listings other than those related to social reasons. We address this concern in two ways. First, in column (3) of Table AVII, we add listing specific fixed effects. In this specification, the identification of the parameters of interest occurs when hosts grow from casual hosts to multi-listing hosts. The coefficients in this specification are similar to the baseline coefficients.

Second, to identify the effect of interacting with the host on reviewing behavior, we compare trips to listings located at same address. Some of these trips are to private rooms when the host is there and others are to entire properties when the host is not there. Table AVIII displays the estimates for specifications with Address fixed effects. Column (1) shows that there is not difference on average between ratings of entire listings and private rooms. Column (2) adds an interaction between listing type and the guest's private recommendation. There is now an important heterogeneity. When guests do not recommend, entire properties are 5 percentage points less likely to receive a high star rating than private rooms. However, when guests do recommend there is no statistically significant difference in high ratings between the two types of listings. In fact, when adding address fixed effects, the effect of social interaction actually increases compared to the baseline specification in Table AVII. We view these results as demonstrating that the room type and the type of host causally affect the willingness of guests to honestly report their experiences in review ratings.

F 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.

We first describe three measures of bias, each with theoretical and practical trade-offs. Our first measure of bias, B_{avg}, is the difference between average experience and the reported experience. The biggest advantage of this measure is that it includes the bias due to sorting into a review. However, of the measures we consider, it requires the most assumptions to calculate. Furthermore, the average can be uninformative if there are multiple sources of bias that push the average review in opposite directions. Our second measure of bias, B_{mis}, is the share of all submitted reviews that are misreported. This measure quantifies the degree of dishonesty in the system. Dishonesty may be important, separately from average bias, because Bayesian updaters can adjust expectations for overall inflation but not for particular instances of lies. The main disadvantage of, B_{mis}, is that it does not measure bias due to sorting into reviewing. Our last 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 (2014)), the platform may especially care about catching these bad agents in the review system.

F.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. Then 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 deviation between the average true experience and the average submitted review is:

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

Where 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.

We assume that, in the absence of retaliation and reciprocity, guests honestly recommend when they leave a review (because the recommendation is anonymous).[22] In order to calculate the empirical analogue to g, we need to make assumptions about selection into reviewing. We first note that the recommendation rate for guests in the incentivized review experiment was lower than in the control. Therefore, in the absence of monetary incentives to review, $Pr(r|g) \neq Pr(r|b)$ and we cannot simply use the rates of recommendations in the data to back out g. Instead, we calibrate g by using the recommendation rates from the incentivized review experiment, which eliminates some of the effect of selection into reviewing. However, because 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.

$$\hat{g} = s_{0,ir,sr}\frac{s_{all,sr}}{s_{0,c,sr}} \tag{9}$$

where $s_{0,ir,sr}$ is the share of positive reviews in the incentivized review (ir) and simultaneous reveal (sr) treatments, $s_{0,c,sr}$ is the share of positive reviews in the ir control and sr treatment, and $s_{all,sr}$ is the share of positive reviews in the entire sr treatment. For $\hat{g}$ to be an unbiased estimate of good experiences, we need to make two more assumptions. First, the rate of positive experiences for those that do not review in the coupon experiment must be equal to the rate of positive experiences in the overall sample. We view this assumption as 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, the relative rate of bias due to sorting must be the same across all types of listings. In the absence of experimental variation, we cannot confirm or reject this proposition. Lastly, we need to measure the conditional review probabilities and mis-reporting rates conditional on leaving a review. We describe how to do so in the next section.

Our second measure of bias is the share of all submitted reviews that are misreported, B_{mis}:

$$B_{mis} = \frac{N_{p|n} + N_{n|p}}{N_{rev}} \tag{10}$$

where $N_{p|n}$ is the number of positive reviews with a negative experience, $N_{n|p}$ is the number of negative reviews with a positive experience, and N_{rev} is the total number of reviews. The practical advantage of this measure is that it requires no assumptions about buyers who do not review for instances that appear in the data.

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

$$B_{neg} = 1 - \frac{N_{n|n}}{N_{all}(1-g)} \tag{11}$$

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

[22] Note, the simultaneous reveal experiment did not affect the average recommendation rates.

G Additional Tables

Table AI: Determinants of Guest Reviews

	Reviewed	
Five Star Rate	0.105***	0.106***
	(0.008)	(0.008)
Past Booker	0.059***	0.059***
	(0.004)	(0.004)
No Reviews	0.026**	0.025*
	(0.013)	(0.013)
No Trips	0.096***	0.098***
	(0.012)	(0.012)
Num. Trips	−0.0004***	−0.0004***
	(0.0001)	(0.0001)
Customer Service	−0.174***	−0.167***
	(0.020)	(0.020)
Entire Property	0.004	0.005
	(0.005)	(0.005)
Multi-Listing Host	−0.095***	−0.084***
	(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	60,579	60,579
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: Experimental Validity Check

Variable	Experiment	Difference	Mean Treatment	Mean Control	P-Value	Stars
Experienced Guest	Simultaneous Reveal	0.00000	0.999	0.999	0.985	
US Guest	Simultaneous Reveal	-0.001	0.283	0.284	0.718	
Prev. Host Bookings	Simultaneous Reveal	-0.167	14.878	15.045	0.259	
US Host	Simultaneous Reveal	0.001	0.263	0.262	0.778	
Multi-Listing Host	Simultaneous Reveal	0.001	0.082	0.081	0.374	
Entire Property	Simultaneous Reveal	-0.001	0.670	0.671	0.840	
Reviewed Listing	Simultaneous Reveal	-0.003	0.764	0.767	0.159	
Observations	Simultaneous Reveal	0.001			0.414	
Experienced Guest	Incentivized Review	0.0005	0.999	0.999	0.163	
US Guest	Incentivized Review	0.001	0.229	0.228	0.904	
Prev. Host Bookings	Incentivized Review	-0.008	0.135	0.143	0.130	
US Host	Incentivized Review	0.0004	0.199	0.199	0.938	
Multi-Listing Host	Incentivized Review	0.001	0.169	0.168	0.760	
Entire Property	Incentivized Review	0.002	0.683	0.681	0.724	
Host Reviews Within 7 Days	Incentivized Review	-0.009	0.736	0.745	0.158	
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 AIII: Effect of Coupon Treatment on Five Star Ratings

	(1)	(2)	(3)	(4)	(5)
Treatment	−0.082***	−0.081***	−0.104**	−0.077***	−0.088***
	(0.010)	(0.009)	(0.048)	(0.009)	(0.017)
Guest Lenient			0.168***		
			(0.057)		
Treatment * Guest Lenient			0.044		
			(0.074)		
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	10,623	10,623	615	10,623	10,623

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 Judiciousness” is a guest specific fixed effect that measure a guest’s propensity to leave negative reviews. Judiciousness is estimated on the set of all reviews in the year proceeding the experiment. 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 AIV: Retaliation and Induced Reciprocity - Host

	Does Not Recommend	Negative Sentiment	Does Not Recommend	Negative Sentiment
	(1)	(2)	(3)	(4)
Treatment	−0.002	0.006	−0.001	0.007
	(0.001)	(0.008)	(0.001)	(0.009)
Guest Low Rating	0.239***	0.312***	0.104***	0.159***
	(0.028)	(0.048)	(0.031)	(0.058)
Guest Review Negative Words			0.348***	0.199*
			(0.093)	(0.119)
Guest Does Not Recommend			0.084	0.157*
			(0.063)	(0.091)
Treatment * Low Rating	−0.177	−0.254***	−0.048	−0.120
	(0.031)	(0.058)	(0.037)	(0.075)
Treatment * Review Negative Words			−0.259***	−0.179
			(0.098)	(0.149)
Treatment * Does Not Recommend			−0.125*	−0.140
			(0.065)	(0.118)
Guest, Trip, and Listing Char.	Yes	Yes	Yes	Yes
Observations	14,384	8,190	10,684	7,520

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

	Reviews First	< 5 Rating (First)	Neg. Sentiment (First)	Neg. Sentiment (First)
	(1)	(2)	(3)	(4)
Treatment	0.0004	0.002	0.007*	0.009**
	(0.002)	(0.004)	(0.004)	(0.005)
< 5 Rating				0.156***
				(0.009)
Not Recommend		0.658***		0.439***
		(0.007)		(0.022)
Treatment * < 5 Rating				−0.008
				(0.012)
Treatment * Not Recommend				−0.024
				(0.031)
Guest, Trip, and Listing Char.	Yes	Yes	Yes	Yes
Observations	38,035	38,033	31,358	29,376

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: 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 AVIII: 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

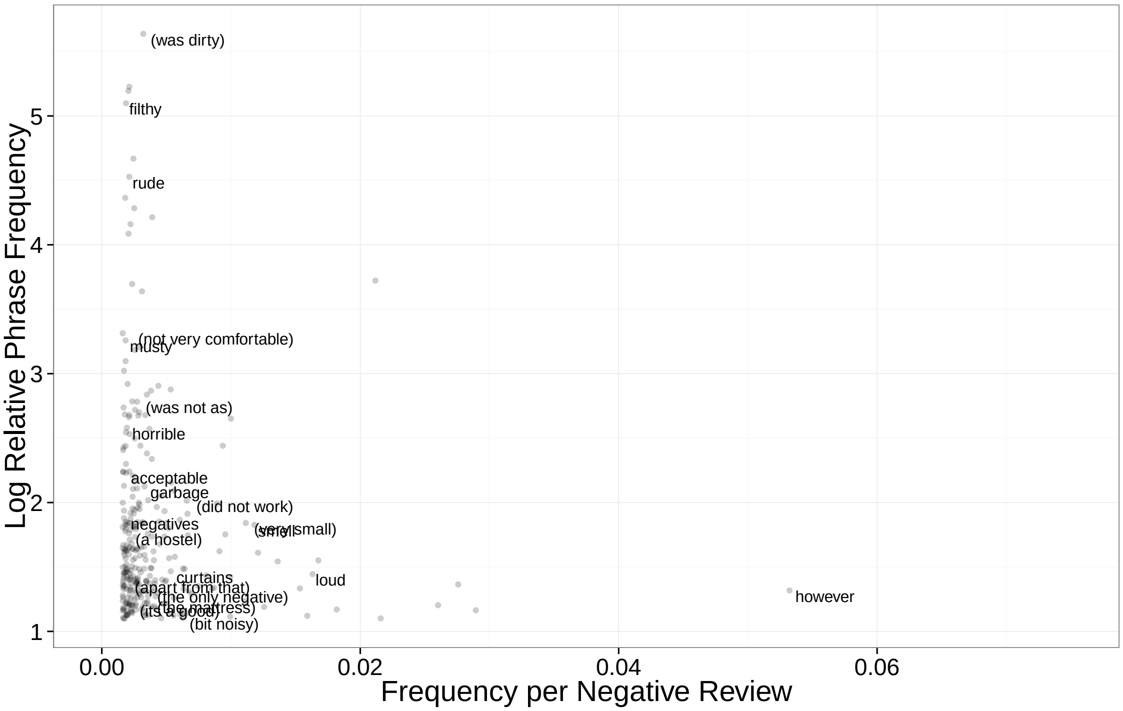
Table AIX: Size of Bias
(Guest does not recommend listing but submits five star rating.)

	Measure of Bias:		
	B_{avg}	B_{mis}	B_{neg}
Counterfactual:	Average	% Misreported	% Negative Missing
Baseline	0.61	0.15	52.02
Simultaneous Reveal	0.58	0.18	49.43
Simultaneous Reveal + No Social Reciprocity	0.43	0.03	46.23
Simultaneous Reveal + No Social Reciprocity + No Sorting	0.03	0.03	31.77
Above + Everyone Reviews	0.03	0.03	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.

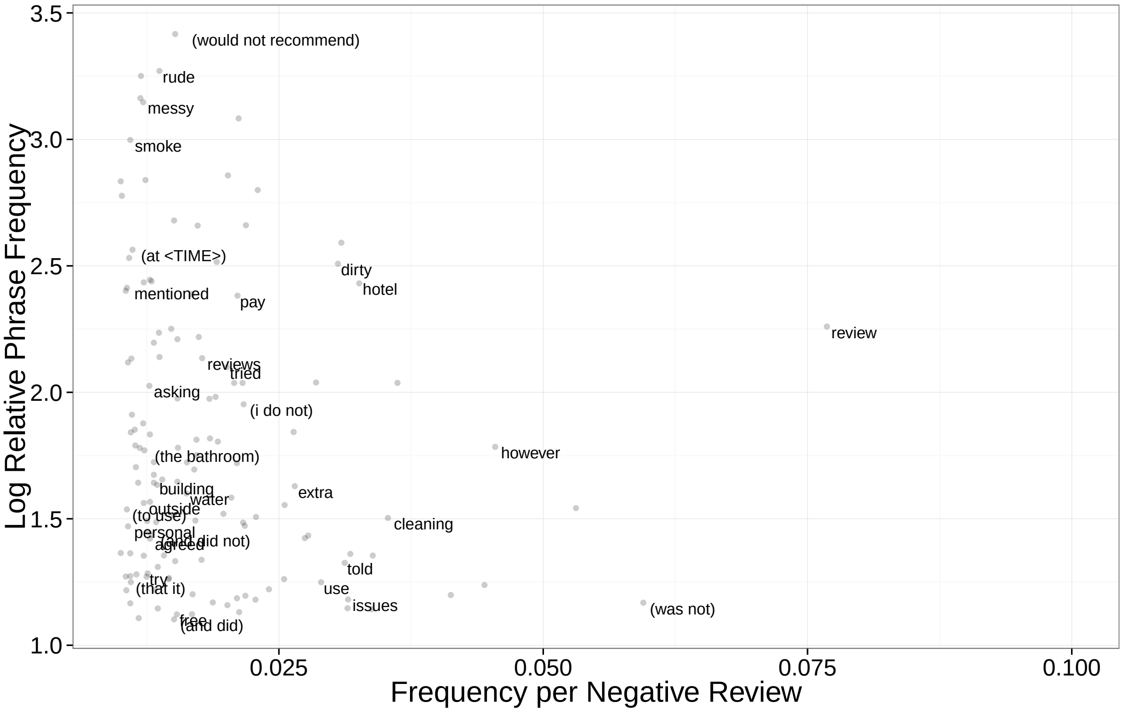
H Additional Figures

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



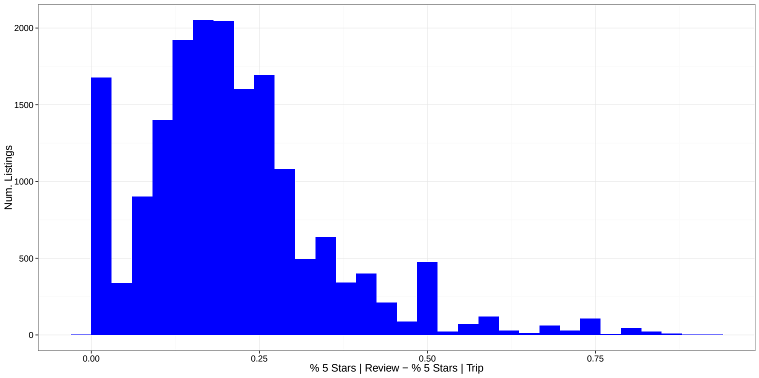
"Relative phrase frequency" refers to the ratio with which the phrase occurs in reviews with a rating of less than five stars.

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



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

Figure A3: 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.