Bias and Reciprocity in Online Reviews: Evidence From Field Experiments on Airbnb (Preliminary - Please Do Not Cite)

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.

December 19, 2014

Abstract

Online reviews and reputation ratings help consumers choose what goods to buy and whom to trade with. However, potential reviewers are not compensated for submitting reviews or making reviews accurate. These conditions can result in review bias, which occurs when some types of experiences are less likely to result in a review or when reviewers misrepresent their experiences in submitted reviews. We study the determinants and size of bias in online reviews by using proprietary data regarding two field experiments on Airbnb. In the first experiment, we induce more consumers to leave reviews by offering them a coupon. Those induced to review report more negative experiences than reviewers in the control group. In our second experiment, we remove the possibility of retaliation and reciprocation in reviews by changing the rules of the review system. We find that fear of retaliation, retaliation against negative reviews, and reciprocity of positive reviews all cause bias but that the magnitude of this bias is smaller than the bias due to sorting. Lastly, we document a new reason for bias in evaluations, *socially induced reciprocity*, which occurs when buyers and sellers interact socially and consequently omit negative information from reviews. This mechanism represents a major challenge for online marketplaces that intermediate transactions involving social interaction.

We are grateful to Jon Levin, Liran Einav, Caroline Hoxby, Shane Greenstein, Mike Luca, Chris Dellarocas, John Horton, Chiara Faronnato, Jeff Naecker, Fred Panier, and participants of the CODE conference for comments. Note: The views expressed in this paper are solely the authors' and do not necessarily reflect the views of Airbnb Inc.

*Primary Author: fradkina@nber.org
†Primary Experiment Designer: elena.grewal@airbnb.com
‡dave.holtz@airbnb.com
§matthew.pearson@airbnb.com

1 Introduction

Online reviews, recommendations, and reputation scores help consumers (and other economic actors) decide what goods to buy and whom to trade with. These reputation systems are especially important for online marketplaces, where economic agents often interact with new trading partners who provide heterogeneous goods and services.[1] However, because reviewers are not compensated for submitting reviews or making them accurate, 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)). Furthermore, as we later show, reviewers may actually have an incentive to misreport the quality of their experiences in reviews. These conditions can result in review bias, which occurs when some types of experiences are less likely to result in a review or when reviewers misrepresent their experiences in submitted reviews. Review bias can consequently lead to worse matches between buyers and sellers (Horton (2014), Nosko and Tadelis (2014)), moral hazard (Hui et al. (2014)), and lower overall levels of trust. Numerous studies have proposed theoretical reasons why bias may occur (Dellarocas and Wood (2007), Bolton et al. (2012)). However, because of data limitations, there has been little evidence about the causes of bias and their relative importance in practice.[2]

In this paper, we use proprietary data regarding two field experiments on Airbnb, a prominent online marketplace for accommodations, to show that some review bias exists and to study its causes. Reputation is particularly important for transactions on Airbnb because guests and hosts interact in person, often in the primary home of the host. Airbnb's review system records a variety of information that helps market participants make decisions about whom to trade with. Both guests (buyers) and hosts (sellers) review each other after the end of a transaction (the checkout date). Each review includes three types of feedback: text, star ratings on a 1 to 5 scale, and anonymous recommendations (which are never displayed on the website). These reviews are generally informative — anonymous recommendations are correlated with public star ratings and guests who call customer support leave lower ratings. However, as in other online marketplaces, reviews on Airbnb are predominantly positive. Over 70% of reviewers leave a 5 star rating and over 97% of reviewers recommend their counterparty.

[1] 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 smaller restaurants.

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

To what extent do these positive reviews reflect honest reports of positive experiences? We use differences in the anonymity of review prompts to measure this bias. Reviewers should be more honest in recommendations than in review text or star ratings because recommendations are anonymous, cannot trigger retaliation, and cannot hurt the person being reviewed. Indeed, we find that over 20% of guests who did not recommend their hosts nonetheless submitted a four or five star rating and omitted negative text from reviews. We then use experiments to quantify the importance of three causes of bias in Airbnb reviews: sorting based on the quality of experience, strategic reciprocity, and socially induced reciprocity. Finally, we use a simple model to show that these causes of bias reduce the rate at which negative experiences are reported in review text by at least 39%.

We proceed by first measuring each cause of bias separately. In section 3 we consider sorting bias, which occurs if those who do not review have worse experiences than those who do review (Dellarocas and Wood (2007) and Nosko and Tadelis (2014)). To test for sorting bias, we conduct an experiment in which we randomly offer guests a $25 coupon to submit a review. Our treatment increases review rates by 6.4 percentage points higher and decreases the share of those reviews that are five stars by 2.1 percentage points. We show that this effect is caused by the fact that those induced to review have worse experiences than the average reviewer.

Sections 4 and 5 contain our results about strategic reciprocity, which refers to a set of behaviors that occur when the second reviewer can respond to the review content of the first review. This can lead to bias if, for example, first reviewers omit negative information from reviews due to the fear of retaliation by the second reviewer. We test for this bias with an experiment that removes the possibility of strategic reviewing responses by hiding any feedback until both the buyer and seller have left a review (or the review window has expired). This experiment precludes a second reviewer from choosing review content in response to the content of the first review. 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 but does not affect recommendation rates.

We use the results of this experiment to show that hosts and guests choose the content of their reviews in response to strategic considerations. We show that guests who review second in the control group sometimes retaliate against negative review text by hosts and reciprocate positive review text. Hosts in the control group also retaliate against negative text by guests but they do not reciprocate positive text. Next, we look at whether the possibility of retaliation changes the behavior of first reviewers. We find that both hosts and guests are more likely to omit negative review text when there is a possibility of retaliation.

Lastly, in section 6, we study socially induced reciprocity, which happens when social

interaction during a transaction induces reviewers to inflate their ratings or review text. Social interaction frequently occurs on Airbnb when guests and hosts send messages and talk to each other while sharing the rented property. Similar interactions also happen in other services marketplaces such as Uber (rides) and Taskrabbit (errands). Conversation 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.[3]

Our strategy for finding the effect of socially induced reciprocity on reviews relies on the fact that different types of trips on Airbnb entail different levels of social interaction. For example, trips to properties with property managers (defined as hosts with more than 3 listings) are less likely to result in social interactions because property managers often manage listings remotely and converse in a transactional as opposed to a social manner. We find that guests who do not recommend the listing are more likely to submit a five star rating for non-property managers than for property managers. Furthermore, reviews of property managers of entire properties are 3.7 percentage points more likely to contain negative text than reviews of casual hosts of private rooms. Our estimates are not diminished after controlling for other measures of the reviewing guest's trip quality (star ratings, anonymous recommendations, and customer service tickets). Therefore, this effect is not caused by differences in the average quality of experiences.

In section 7, we describe our methodology for inferring the frequency of negative experiences and the rate at which they are reported. We then propose and evaluate several measures of bias. Our first measure of bias is the difference between the share of reviews with negative text and the share of negative experiences for all trips. Even when all biases operate, we find that the average bias is only 1.7 percentage points. Similarly, the share of reviews in which the text does not reflect the recommendation of the guest, comprises fewer than 1% of reviews. Therefore, assuming that guests reply honestly in the anonymous recommendation prompt, most positive reviews accurately reflect a positive experience.

We then study whether negative experiences are reported in reviews, if they do occur. We find that 68% of all negative trips are not accompanied by negative review text when all

[3]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!"

biases operate. This number falls to at most 41% when the three biases we study are removed. The remaining bias is caused mostly be the fact that, even without sorting, approximately 30% of guests do not review and therefore do not report negative experiences. Therefore, although negative experiences are rare, the text of Airbnb reviews frequently does not reflect those negative experiences. With regard to the relative importance of the causes of bias, we find that removing bias from socially induced reciprocity and sorting each result in a reduction of at least 10 percentage points in the share of missing negative reviews. On the other hand, the decrease in bias from removing strategic reciprocity is only 5 percentage points.

Our results broadly inform reputation system design in online platforms. Platforms must choose what content to display (and how to elicit it), whether the content displayed is linked to the reviewer or aggregated, and when the content is publicly revealed on the website. Each design choice has theoretical trade-offs, many of which involve the degree of bias in the review system. In addition to directly informing users, reviews are also important because they interact with other market design choices such as search ranking algorithms (which often use reviews as features) and certifications (e.g. Ebay's Trusted Seller or Airbnb's Superhost). For example, Nosko and Tadelis (2014) show that Ebay's search algorithms create better matches when they account for review bias using a seller's Effective Positive Percentage (EPP),[4] the ratio of positive reviews to transactions (rather than total reviews). We provide the first direct evidence that buyers who don't 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. Relatedly, Bolton et al. (2012) propose a "blind" review system to remove bias from strategic reciprocity and evaluate it in laboratory experiments. We are the first to study the effects of such a review system in a field experiment. Although we find that this intervention reduces bias, non-strategic sources of bias are relatively more important in our setting.

Lastly, our findings relate to a large behavioral economics literature focusing on giving and reciprocity. Numerous laboratory studies have found that giving decreases with social distance (Bohnet and Frey (1999)) and increases with non-binding communication (Sally

[4] A related literature attempts to infer the "true" quality of a seller from observed review and transaction data. Dai et al. (2012) and Dellarocas and Wood (2007) propose structural econometric approaches to de-bias public reviews (making particular assumptions about the types of bias in reputation systems).

(1995)). 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. We show that these laboratory results carry over to reviewing behavior.

Another related literature shows that participation in giving games is actually an endogenous variable (Malmendier et al. (2014), Lazear et al. (2012), DellaVigna et al. (2012)). These papers 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 5 star rating,[5] and private comments to the host (shown in Figure 2). 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 subscores 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. Finally, the guest can provide private feedback directly to Airbnb about the quality of the trip using a text box and a "likelihood to recommend" (LTR) question prompt.[6]

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 can also leave a written review of the guest 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 115,000 trips between May 10, 2014 and June 12, 2014, which are in the control group of the subsequent experiments.[7] The summary statistics for these trips are shown in columns 1 and 2 of Table 1. Turning first to review rates, 65% of trips result in a guest review and 72% result in a host review.[8] 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. The fact that hosts review at higher rates and review first is due to two facts. First, because hosts receive inquiries from other guests, they check the Airbnb website more frequently than guests. Second, hosts have more to gain than guests from inducing a positive review by a guest and therefore tend to review first.

We first consider guest reviews of hosts. 97% of guests who submit a review recommend a listing in a question prompt that is marked as anonymous and whose answers are not displayed anywhere on the website. This suggests that most guests do have a good experience.

[5] 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.

[6] The prompt for the LTR is: "On a scale of 0-10, how likely would you be to recommend Airbnb to a friend or colleague?". This question is frequently used in industry to calculate the "Net Promoter Score".

[7] 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.

[8] These review rates are similar to the review rate on Ebay (65%) and smaller than the review rate by freelancers on Odesk (92%). However, review rates in peer-to-peer online marketplaces are much higher than review rates on more traditional online retailers such as Amazon.com.

Figure 4 shows the distribution of star ratings for submitted reviews. Guests submit a five star overall rating 74% of the time and a four star rating 20% of the time. There is no spike in the distribution for 1 star reviews, as seen on retail sites like Amazon.com.[9] Guest reviews on Airbnb are overwhelmingly positive compared to other hotel review websites. For example, the rate of five star reviews is 31% on TripAdvisor and 44% on Expedia (Mayzlin et al. (2014)). It is unlikely that Airbnb stays are 30% more likely to result in a very positive experience than hotel stays. The high rate of five star reviews is suggestive of at least some review inflation on Airbnb. However, differences between reviews on Airbnb and Expedia might be due to reasons unrelated to quality, including the fact that different types of individuals leave reviews on the two websites, that different types of establishments operate on the two websites, and that Airbnb's review system is two-sided. Airbnb also prompts guests to rate their stays with category star ratings. These ratings are more informative than the overall rating, with fifty percent of trips having at least one sub-rating that is lower than five stars.

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)). Figure 5 shows 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 ("was dirty"), smell ("musty"), unsuitable furniture ("the mattress"), sound ("loud") and sentiment ("did not work", "was not as", "however").

Recall that guests are asked for three types of feedback about their stay: public review text, star ratings (which are displayed as averages and are not linked to the reviewer), and the recommendation (which is anonymous and not displayed on the website). What is the relationship between these ratings? If guests report their experiences honestly, then there should be little difference between these types of feedback for a given review. However, guests' answers to these questions differ greatly in a given review.

Figure 4 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 4 or 5 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. However, the fact that fewer than 5% of reviewers recommend a listing when they submit a lower than four star rating suggests that

[9]Review rates are much lower for retail websites.

guests indeed understand the review prompt. Another concern is guests may interpret the review prompt in many ways. 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.

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 5. 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 section 5.

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 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 are respectful and do not inconvenience their hosts.

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 turn to the hosts' review text to determine the characteristics of a bad guest. Figure 6 shows phrases that were at least 4 times as likely to occur in reviews in which the host does not recommend the guest. Negative reviews contain phrases concerning sentiment ("bad", "would not recommend"), personal communication ("rude", "complained"), money ("pay", "money"), and cleanliness or damage ("smoke", "mess", "damage", "issue"). This text shows that bad guests complain to their hosts (often about money), do not follow rules, are unusually dirty, or break something. We study the determinants of a host's decision to omit information about guests in section 5.

3 Sorting Bias and the Incentivized Review Experiment

In this section we study the causes and magnitude of sorting bias on Airbnb. This bias occurs when the quality of a user's experience influences the probability of a review (Dellarocas and Wood (2007)). For example, if potential reviewers value reviewing positive experiences more than reviewing negative experiences, then the observed reviews on the website would be upwardly biased. This bias can manifest itself in two ways, each with differing implications for consumer welfare. First, there might be aggregate bias, which occurs when reviews overstate the true quality of each listing on average. Aggregate bias might be accounted for by Bayesian consumers. That is, a potential guest might know that a listing with high ratings is still likely to have had some unsatisfied customers and would value the listing accordingly. Second, there might be bias at an individual level if bias differs amongst listings with similar ratings (or other observable characteristics). In those cases, even knowledgeable guests may mistakenly book a low quality listing because that listing has especially biased reviews.

To demonstrate that there is potential heterogeneity in bias, we plot the distribution of two quality measures: the share of five star reviews out of all reviews and the Effective Positive Percentage (EPP)[10] proposed by Nosko and Tadelis (2014). The first measure is potentially biased by non-reviews whereas the second is not. Figure 11 plots these measures for the sample of highly rated listings (> 4.75 average overall star rating) and with at least 3 reviews.[11] This sample is chosen because Airbnb only displays star ratings after 3 reviews are submitted and rounds the star rating the nearest .5 stars. Therefore, the listings in this sample seem the same to guests on the overall star rating dimension.

There are three interesting features in the ratings distributions. First, the EPP is approximately 25% lower on average than the share of five stars. Second, there are more than 6000 fewer listings with a perfect 1 rating as measured by EPP than by the share of five star reviews. Lastly, there is more dispersion in the EPP measure than in the share of five star reviews. Figure 8 plots the difference between the share of five star ratings and the EPP. The interquartile range for this difference is 16% - 27%. We view this as evidence of a potentially large heterogeneity in bias between listings. The extent of that bias depends on the correlation between a non-review and the quality of the guest's experience.

In the next two sections we document that non-reviews are indicative of worse experiences and show how incentivizing people to leave reviews can alleviate this bias. Our empirical strategy is twofold. We first show that the quality of the listing, the experience during the trip, and guest characteristics all influence the probability that a guest leaves a review. We

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

then describe and evaluate an experiment that induces guests to leave reviews.

3.1 Determinants of Reviewing

Table 2 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.

3.2 The Incentivized Review Experiment

Airbnb guests rely on reviews to decide on whether to book a particular listing. However, new listings do not have reviews and are therefore at a competitive disadvantage to positively reviewed listings. Each trip provides a potential review to the listing but not all trips results in reviews. In April 2014, Airbnb began an experimental program intended to help non-reviewed listings and to learn about the experiences of guests who do not review. This program worked as follows. Trips to non-reviewed listings, for which the guest did not leave a review within 9 days were assigned to either a treatment group or a control group (each assigned with a 50% probability at a host level). Guests in the treatment group received an email offering a $25 Airbnb coupon while guests in the control group received a normal reminder email (shown in Figures 9 and 10).

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. We therefore limit our analysis to the first trip to

a listing in the experiment. Table 3 displays the balance of observable characteristics in the 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.

Table 4 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 11 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. The first mechanism is that there could be a sorting bias where those guests who do not review have worse experiences conditional on observables. The second reason is that the guests who are induced to review by the experiment are different in their judiciousness than the guests who do review. Lastly, the fact that Airbnb offered a coupon and reminded guests of their responsibility to review might have induced guests to be more judicious. We test for these alternative explanations in Table 5. Column (1) 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.[12] 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.

Lastly, we test whether the delayed review timing by guests in the experiment is driven by fear of host retaliation. Suppose that guests omit reviews because they are afraid of retaliation and that the coupon induces those guests to return to the site to leave a review. If fear of retaliation affects the review ratings, then we would expect the ratings of guests in the treatment to be especially low for cases when the host has already left a review. To test for this mechanism, we add controls for whether the host reviewed the guest first in Column (5). We find that guest ratings are actually higher when the host reviews first. Furthermore, those induced to review by the treatment are not more likely to leave negative ratings if the host has already reviewed. Therefore, the fear of retaliation is not driving those affected by the coupon to omit reviews.

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

$$e = \frac{s_{\leq 9} r_{\leq 9} + (s_{ctr} + t_{rev})(r_{ctr} + t_r)}{s_{\leq 9} + s_{ctr} + t_{rev}} - \frac{s_{\leq 9} r_{\leq 9} + s_{ctr} r_{ctr}}{s_{\leq 9} + s_{ctr}} \tag{1}$$

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. Furthermore, "$\leq$ 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 6 uses 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,

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

it would still have an effect by inducing more of these guests to review. The experiment decreases the overall share of 5 star ratings by 2.1 percentage points and the share of reviews with recommendations by .5 percentage points. These effects don't 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. In section 7, we use the results of this experiment to impute the distribution of ratings if every guest was forced to review.

4 The Simultaneous Reveal Experiment

Our second experiment changes 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. Any submitted review was immediately posted to the website. This setup allowed for the second reviewer to see the first review and to retaliate or reciprocate in response. Since retaliation and reciprocation reflect strategic concerns rather than the quality of the experience, this mechanism was causing review bias. To remove this source of bias, we changed when reviews were revealed and tested this change in an experiment.

The experiment consists of two treatments and a true control, each assigned with equal probability to all listings on the website (with the exception of a 5% holdout group that is excluded from all experiments). The first treatment changes the potential time to review to 14 days for both guests and hosts. The second "simultaneous reveal" treatment hides reviews until one of two conditions holds: the other party submits a reviews or 14 days since the checkout date pass. The review window was modified because, otherwise, some reviews in the "simultaneous reveal" group might have been hidden for an entire month. For the rest of this paper we refer to the "short expiration" treatment as the control and the "simultaneous reveal" treatment as the treatment.

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. Table 3 shows comparisons of key observable variables between the treatment and control groups. There is no statistically significant difference between guest or listing characteristics in the two groups. 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 ignore this difference because we find balance on all observables and the overall difference in observations is tiny.

Table 1 shows the summary statistics for the treatment and control groups in the "simultaneous reveal" experiment. The treatment increases review rates for guests by 2 percentage

points and for hosts by 7 percentage points. The rate of five star reviews by guests decreases by 1.6 percentage points, while the recommendation rate decreases by .4 percentage points. Furthermore, the anonymous recommendation responses by hosts stay at 99% of all reviews. However, the text of the submitted reviews does change. The rate of negative sentiment (calculated using the methodology described in section 2) in guest reviews of hosts increases from 16% to 18%. This suggests that the experiment did have the intended effect of allowing people to be more honest in their public feedback. However, the overall size of the effect on public ratings is small. *Private feedback* increases more than public feedback, with a 6 percentage point increase for guests and a 2 percentage point increase for hosts. Lastly, hosts in the experiment review more quickly and are 3 percentage point more likely to review first in the treatment group. In the next section we explore potential explanations for these effects.

5 Strategic Reasons to Review

In this section, we use experimental variation to quantify the importance of strategic reviewing on Airbnb. Strategic motivations influence a reviewer to leave a particular type of review in anticipation of future reviews by the counter-party or in response to an earlier review. There are two types of strategic motivations, those of the first reviewer and those of the second reviewer. The first reviewer might be influenced to leave a more positive review than his true experience for one of two reasons. First, the reviewer may be afraid of retaliation. Second, the reviewer may want to induce the second reviewer to leave a positive review. In turn, the second reviewer might be influenced by a negative first review to retaliate by leaving a negative review even if her experience was positive. Alternatively, the second reviewer might be influenced by a positive first review to leave a positive review even if her experience was negative.

5.1 Evidence for Retailiation and Reciprocity

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 estimating equation is:

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

where y_{gl} is a 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 there is positive reciprocity, then the guest should consequently submit a more positive review then if he had not seen the first review. This response corresponds to α_0 being positive when y_{gl} is an indicator of a negative experience. Second, if there is retaliation against negative host reviews, we would expect α_2 to be positive and α_4 to be negative. 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. Here, 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.

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.

Table 7 displays estimates of Equation 2 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 overall treatment effect 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 6 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.[13] 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.

Next, we consider the same specification for cases when hosts review second. Table 8 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 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.

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. To do so, note that 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 these incentives, 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:

$$y_{gl} = \alpha_0 t_l + \alpha_1 DNR_{gl} + \alpha_2 DNR_{gl} * t_l + \epsilon_{gl} \tag{3}$$

[13]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.

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 and α_2 to be positive because first reviews should be more honest in the treatment, and because those that do not recommend should be even more likely to have negative comments.

Table 9 displays estimates of Equation 3 for first reviews by hosts. Column (1) displays the effect of the treatment on the probability that a host reviews first. Hosts are 2.8 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. Column (2) shows the effect of the treatment on the recommendation rates of hosts. There is a small but statistically significant decrease in the rate of recommendations by hosts, likely reflecting selection into reviewing first. Columns (3) and (4) 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 points increase in the overall rate of negative text in first host reviews. However, column (4) shows that this effect is completely 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. This demonstrates that hosts are aware of strategic considerations and omit negative feedback from public reviews even if they have a negative experience. Furthermore, this is a large effect, given that hosts omit negative feedback over 50% of the time when they do not recommend.

We run the same set of specifications for guests' first reviews in Table 10. 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 11 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.

6 Socially Induced Reciprocity

Ratings on Airbnb remain high when compared to Expedia, even when there is no possibility of retaliation or induced reciprocity (see Table 1). In this section, we document that socially induced reciprocity is one reason for these high ratings. Socially induced reciprocity occurs when buyers and sellers socially interact with each other and consequently omit negative feedback.

Stays on Airbnb frequently involve social communication between guests and host. Guests typically communicate with hosts about the availability of the room and the details of the check-in. Furthermore, guests and hosts often socialize while the stay is happening. Unplanned 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. Previous experimental studies show that social communication can affect reviewing behavior for a variety of reasons including empathy (Andreoni and Rao (2011)), social pressure (Malmendier et al. (2014)), and the increased potential for repeated interactions.

Internal Airbnb surveys have asked guests why they do not submit (negative) reviews, even in the simultaneous reveal treatment. Respondents frequently mentioned reasons that are directly related to the empathy generated during social interactions. First, 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-property manager (defined as hosts with more than 3 listings). Property managers 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. We focus on guest reviews in this section because host negative reviews are so rare and because we have fewer measures of hosts' ratings of guests.

Figure 13 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 5 star overall rating than guests staying with property managers. That is, even though all guests in the sample would not recommend the listing they stayed at, those staying with property managers were more likely to voice that opinion publicly.

Our regression specification to formally test for this effect is:

$$y_{gl} = \alpha_0 PR_l + \alpha_1 PM_l + \alpha_2 R_{gl} + \beta' X_{gl} + \epsilon_{gl} \tag{4}$$

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 property manager, 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 property managers induce less reciprocity in guests.

Tables 12, 13, and 14 display the results of the above specification for overall ratings scores and negative sentiment. Turning first to the effect of socially induced reciprocity in ratings, Column (1) of 12 shows the differences in ratings between different types of trips controlling for the guest recommendation. We find that trips to private rooms have a .03 higher star rating and trips to property managers have a .09 lower star rating. Furthermore, the difference in ratings between trips to property managers and trips to casual hosts is twice

as large when the guest does not recommend. Columns (2) and (3) also add controls for the guest's private feedback to Airbnb and the guest's likelihood to recommend Airbnb. The coefficients are similar across specifications.

One worry with the above exercise is that guests who stay at private rooms with casual hosts may be different from other guests. This difference may lead those guests have different reviewing styles, holding all else equal. To account for this possibility, we estimate Equation 4 with guest specific fixed effects. However, because we need multiple observations per guest, we expand our sample to include all trips with a checkout between January 1, 2014 and November 1, 2014. The estimates from this specification are show in Table 13. Column (1) displays our baseline specification while Column (2) controls for the guests answer to the "Likelihood to Recommend Airbnb" prompt. Both effects of interest, for the private room and the property manager still exist when controlling for guest fixed effects. The effect of being a property managers increases in magnitude from .093 to .112 stars on average. However, the magnitude of the effect of a private room shrinks from .025 to .005. This demonstrates that guests who choose to stay in private room review in different ways than guests who stay in entire properties.

Next, we turn to the effect of socially induced reciprocity on the sentiment of review text. Table 14, columns (2) - (4) display the coefficients on property manager and private room, with progressively more controls. Column (2) 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 property manager are 2.2 percentage points more likely to leave negative text. The effect of ratings on sentiment is in the expected direction, with 5 star ratings being 50 percentage points less likely to contain negative text. Column (3) adds additional controls for the lowest sub-category rating that a guest submits. The coefficients on room type and property manager 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 (4) 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 property manager increases to 2.8 percentage points.

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 property manager. Furthermore, because trips to entire properties with property managers often have social interaction, this estimate is likely to be an underestimate of the true effect of socially

induced reciprocity.

7 How Large is the 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.

7.1 Empirical Analogues of Bias Measures

Suppose that each trip results in a positive experience with probability, g, and a negative experience (denoted n) with probability, 1 - g. 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:

$$\bar{s} = \frac{gPr(r|g)Pr(s_g|g,r) + (1-g)Pr(r|n)Pr(s_g|n,r)}{Pr(r)} \tag{5}$$

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

$$B_{avg} = (1-g)\frac{Pr(r|n)Pr(s_g|n,r)}{Pr(r)} - g(1-\frac{Pr(r|g)Pr(s_g|g,r)}{Pr(r)}) \tag{6}$$

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).[14] 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.

$$\hat{g} = s_{0,ir,sr}\frac{s_{all,sr}}{s_{0,c,sr}} \tag{7}$$

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}:

$$B_{mis} = \frac{N_{p|n} + N_{n|p}}{N_{rev}} \tag{8}$$

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:

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

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.

7.2 The Size of Bias

The goal of the exercise in this section is to quantify the degree of bias caused by each mechanism discussed in this paper. We use one specific measure of bias: when a reviewer does not recommend the reviewee but leaves no negative textual feedback (i.e. s_g corresponds to positive textual feedback). 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 retaliate 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 four scenarios, each with progressively less bias. Scenario 1 is one in which all three biases: sorting, strategic, and social operate. This corresponds to the control group in the simultaneous reveal experiment. Scenario 2 removes the strategic bias and corresponds to the treatment group of the simultaneous reveal experiment. In both of these cases, we can calculate the components 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 and $\widehat{Pr(r|n)} = \frac{\widehat{Pr(n|r)} * \widehat{P(r)}}{(1-\hat{g})}$, where the probabilities of non-recommendations and reviews are observable in the data. Scenario 3 further removes

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

social bias in the reviewing process. To do so, we let $Pr(\widehat{s_g|n,r})$ equal to this rate just for stays with property managers in entire properties. This change shifts the probability of mis-reporting a non-recommendation from 68% to 54%. Lastly, scenario 4 removes sorting bias from reviews. This is operationalized by replacing the share of all reviews that don't recommend the listing from .087 (its rate in the data), to $1 - \hat{g} = .105$. Note, the no-sorting calculation still keeps the overall review rate equal to the review rate in the simultaneous reveal treatment.

Table 15 displays each measure of bias for all 4 scenarios described above. We first turn to the case when all biases are present (row 1). In this scenario, positive reviews occur 1.7% more of the time than positive experiences. Furthermore, 1.4% of all reviews mis-represent the quality of a guests experience and 68% of negative experiences are not reported in text. Removing strategic considerations changes these numbers by less than .05 in all cases. The small aggregate effect of strategic motivations is due to the fact that, while the simultaneous reveal treatment did reduce positive reviews for guests who recommended, it had no additional effect on guests who did not recommend. Therefore, we conclude that strategic motivations have little effect on truly negative reviews on Airbnb.

Row 3 shows the bias in the case where social reciprocity is removed as a motivation for reviews. The overall bias is now .9%, while the share of misreported reviews is .8% of all reviews. This represents a drop in bias that is an order of magnitude larger than the drop in bias when strategic motivations are removed. Furthermore, since there is still likely to be social bias for property managers with entire properties, our results are an underestimate of the true effect of social bias.

In row 4, we remove sorting bias. The average bias falls an additional .5 percentage points and the share of negative experiences missing drops to 41% due to the fact that a higher percentage of those with negative experiences now review. 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. 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 some of the additional reviewers in this scenario have negative experiences.

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, negative experiences are often not reported in review text on the website. If the above biases were eliminated, then at least 27% 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 simultaneous 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.

This paper is a first step in a comprehensive analysis of the effects and design of reputation systems. Our future work will study how the review bias that we document affects equilibrium outcomes such as transaction volume and welfare.

References

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

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

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

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.

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.

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.

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.

9 Figures

Figure 1: Reviews on Listing Page

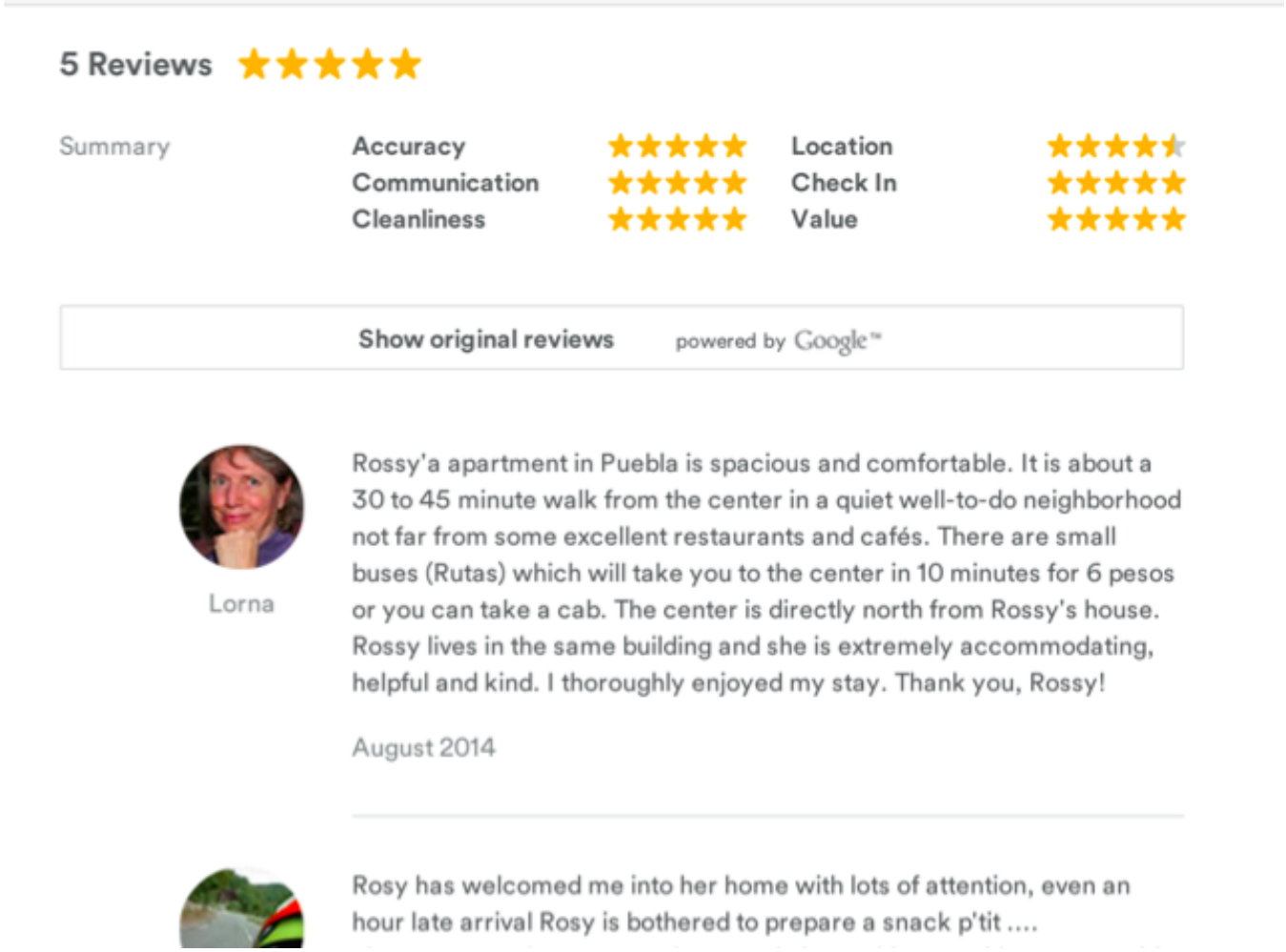


Figure 2: Review of Listing Form (Page 1)

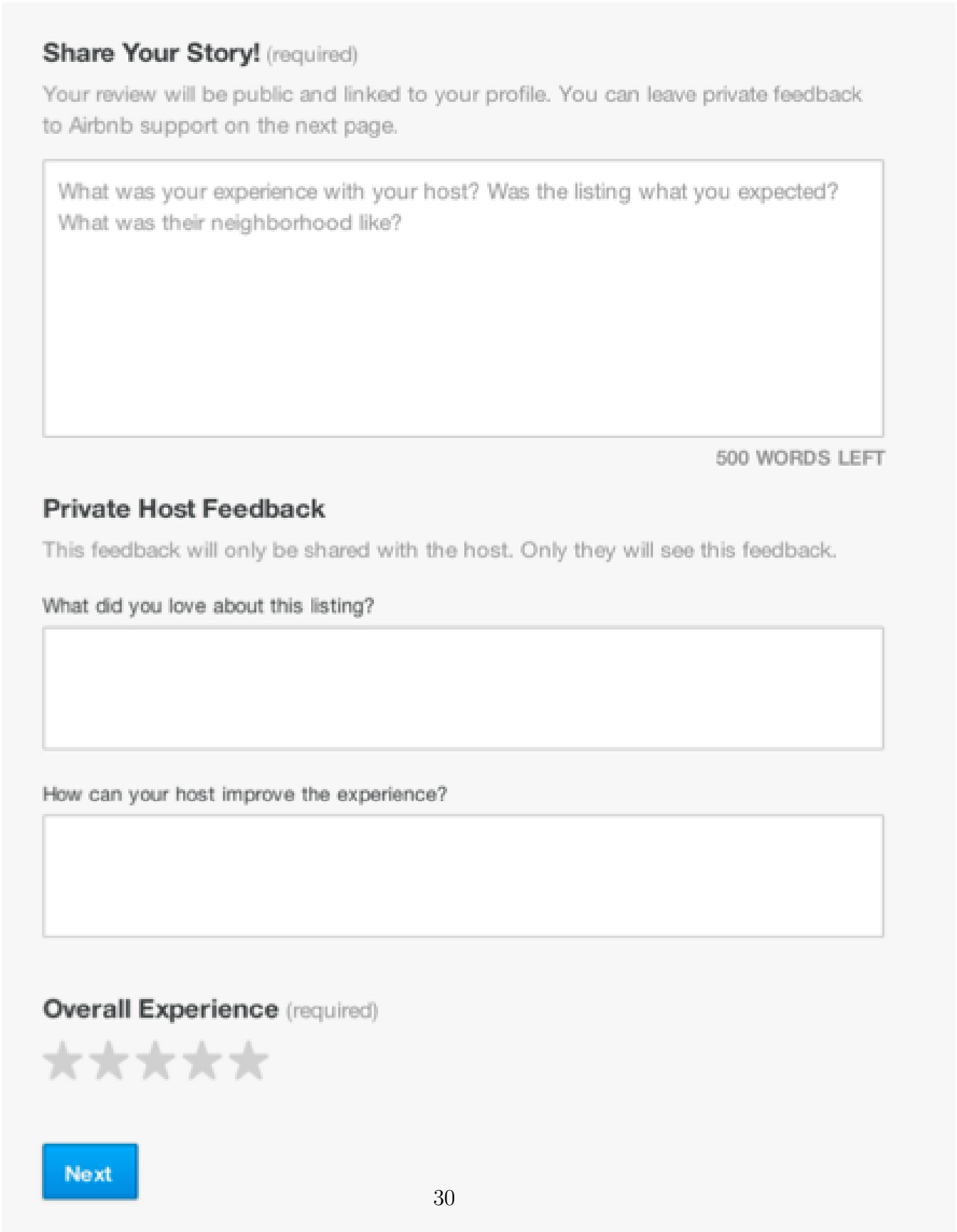


Figure 3: Review of Guest Form (Bottom of Page 1)

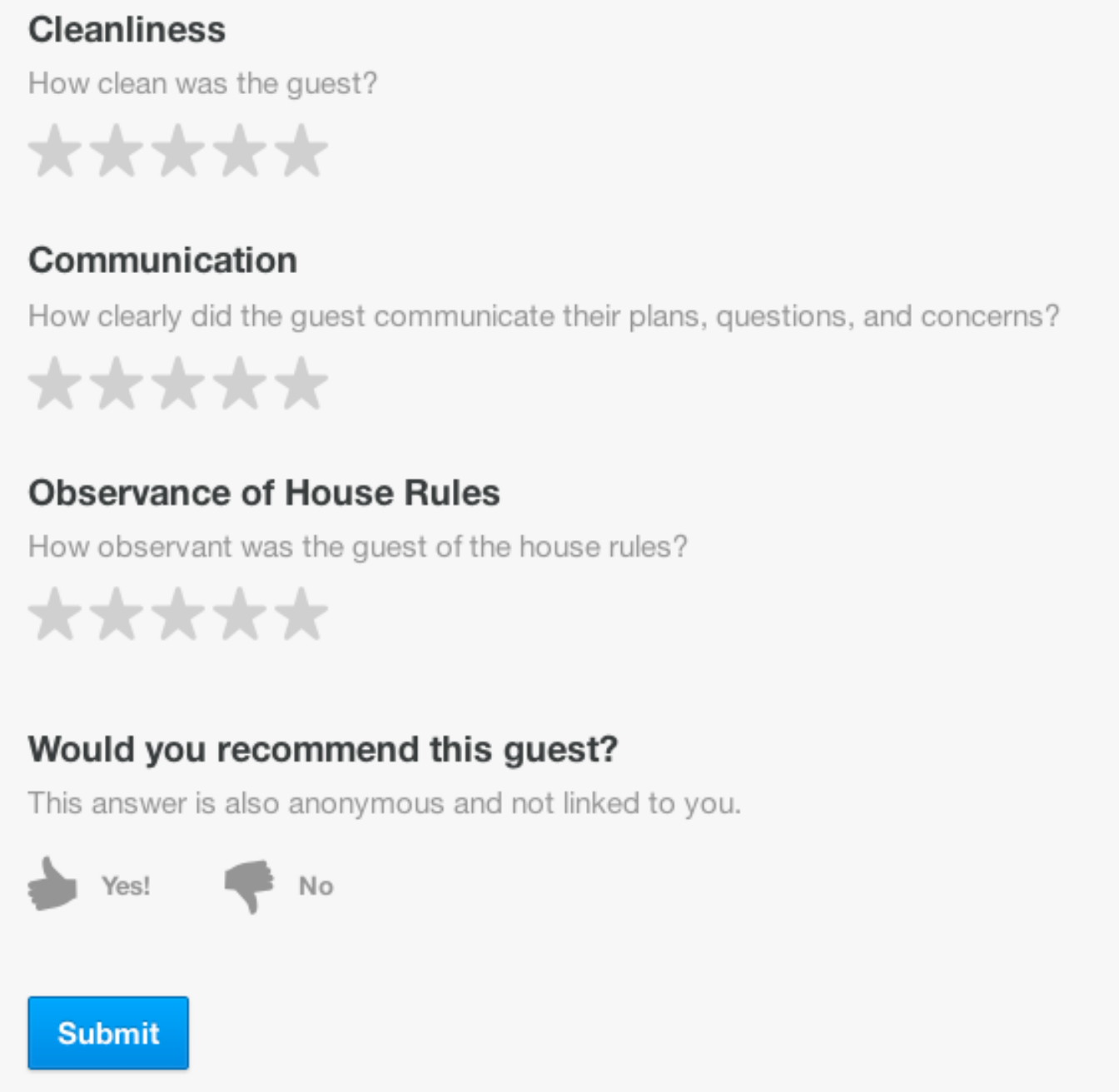
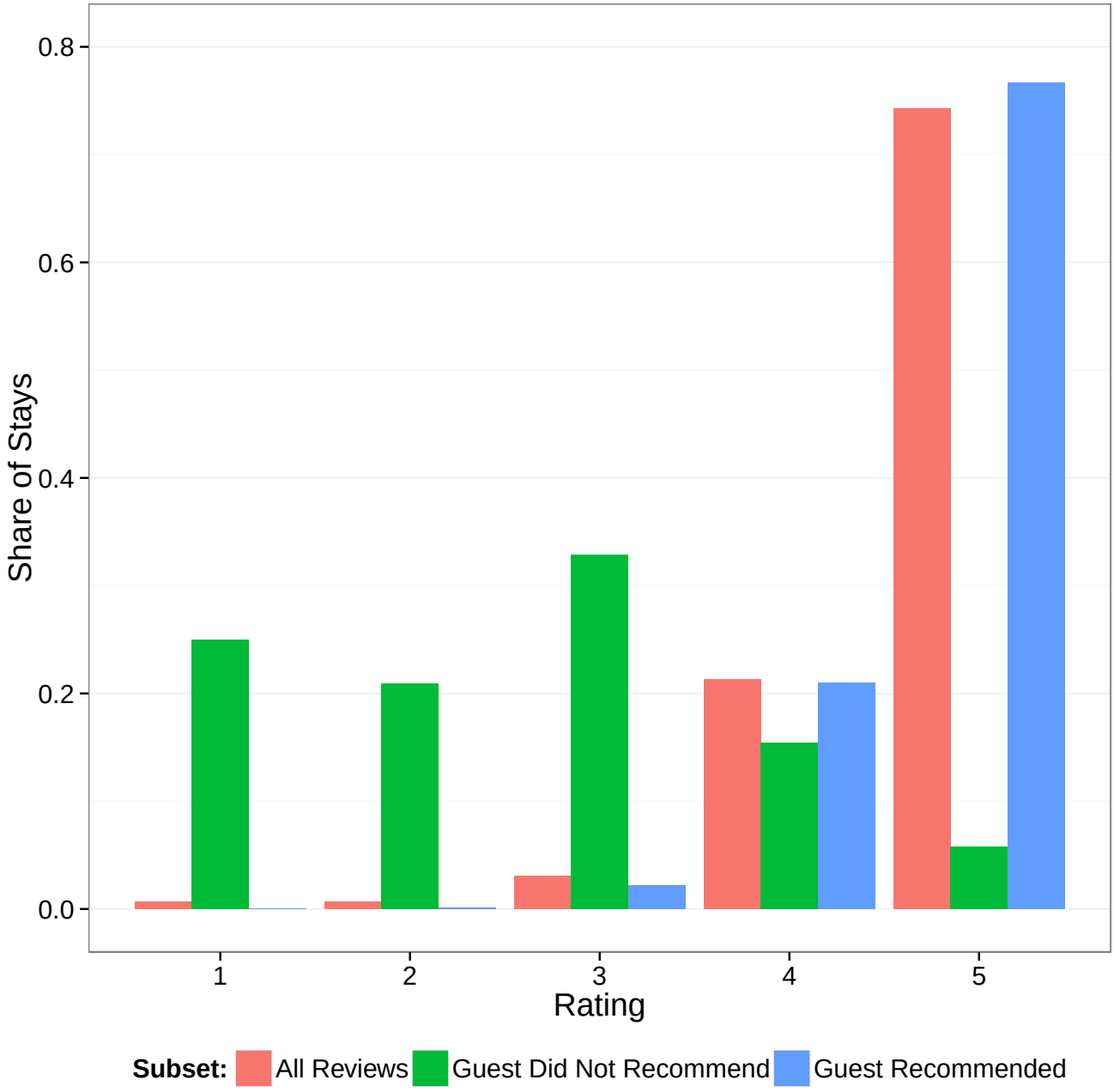
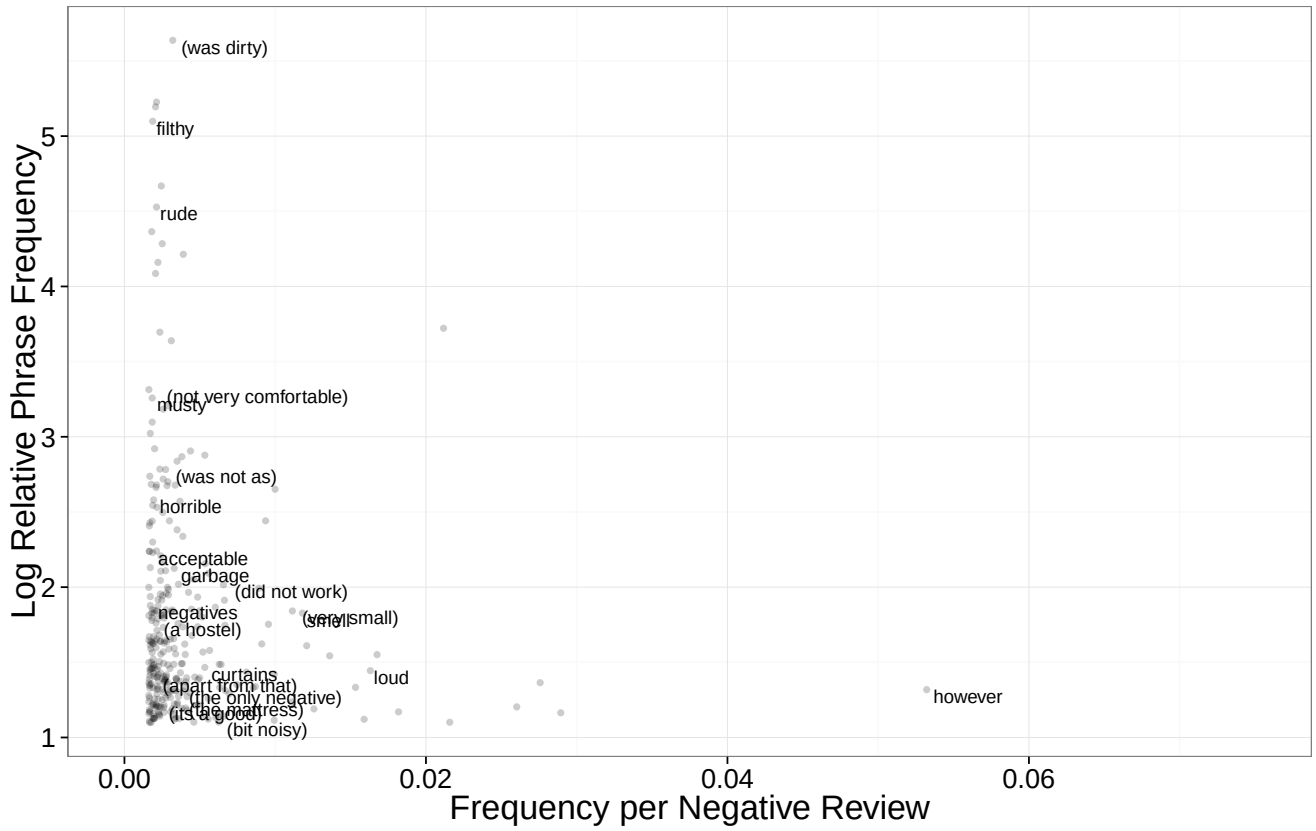


Figure 4: Distribution of Guest Overall Ratings of Listings



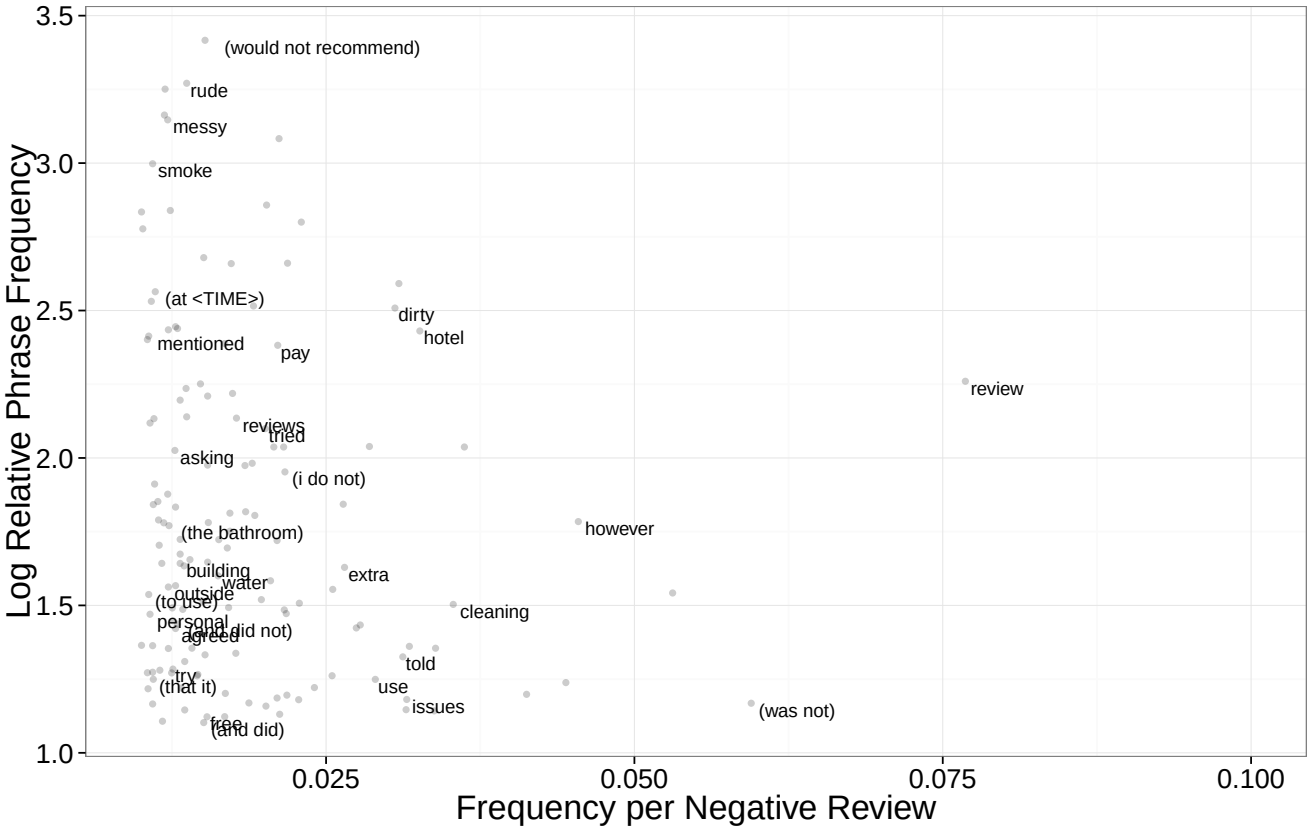
The above figure displays the distribution of submitted ratings in the control group of the simultaneous reveal experiment. Only first stays for each listing in the experimental time period are included. "Guest Did Not Recommend" refers to the subsample where the guest stated that they would not recommend the listing in anonymous prompt. "Guest Recommended" is the analogous sample for those that did recommend the listing.

Figure 5: Phrases Common in Low-rated Reviews by Guests



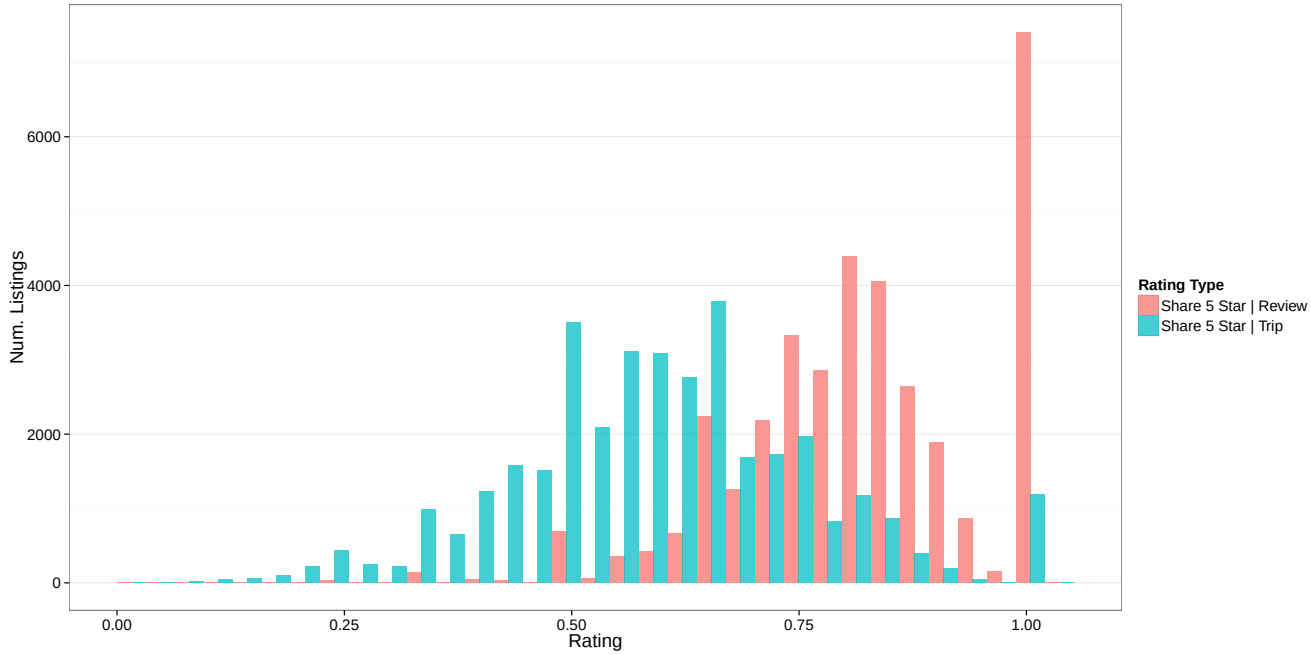
This figure displays all phrases (1, 2, and 3 words) which were at least 4 times as likely to appear in reviews where the guest left a lower than 5 star overall rating than in reviews where the guest left a 5 star rating. In order to be included in this measure, each phrase must have appeared at least 50 times in the set of reviews with a lower than 5 star rating.

Figure 6: Phrases Common in Low-rated Reviews by Hosts



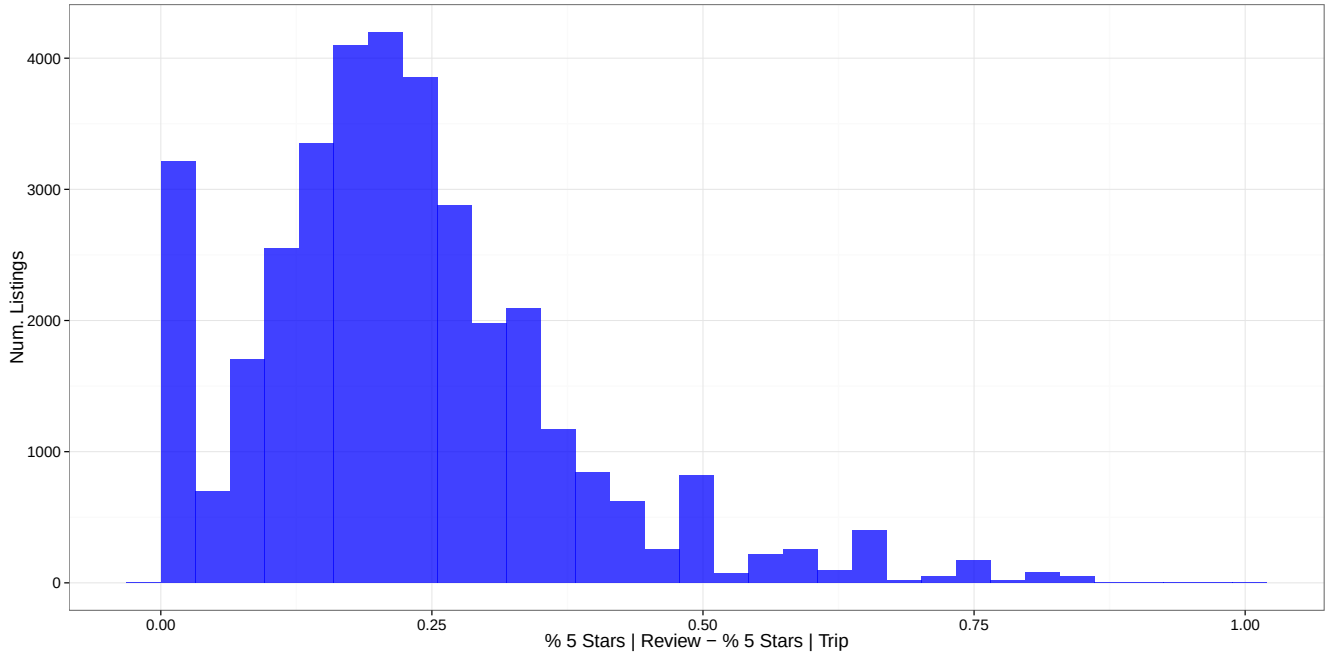
This figure displays all phrases (1, 2, and 3 words) which were at least 4 times as likely to appear in reviews where the host did not recommend the guest than in reviews where the host did recommend the guest. In order to be included in this measure, each phrase must have appeared at least 1% of the time (or 188 times) in reviews where the host did not recommend the guest.

Figure 7: Histogram of Ratings per Listing



This figure displays the distribution of rating at a listing level for two measures of ratings. The first measure is the share of reviews for a listing that have an overall 5 star rating. The second measure is the share of trips (whose review period has ended). The sample includes listings with at least 3 trips and at least a 4.75 average rating in the sample. Because Airbnb rounds star ratings, all listings in the sample are shown as having a 5 out of 5 star rating on the site.

Figure 8: Histogram of Difference in Ratings per Listing



This figure displays the distribution of the difference between the share of five star reviews out of all reviews and the share of 5 star reviews out of all trips. The sample includes listings with at least 3 trips and at least a 4.75 average rating in the sample. Because Airbnb rounds star ratings, all listings in the sample are shown as having a 5 out of 5 star rating on the site.

Figure 9: Coupon Experiment - Treatment Email

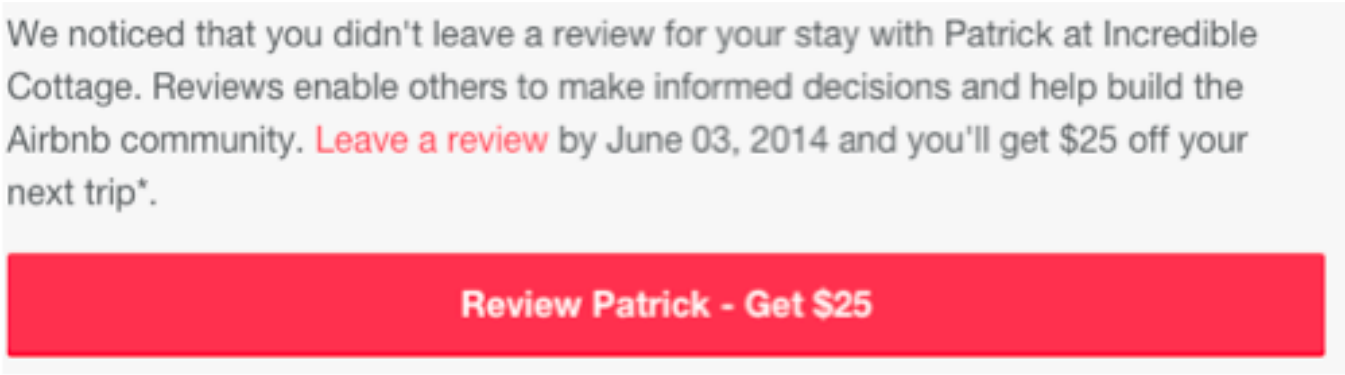


Figure 10: Coupon Experiment - Control Email

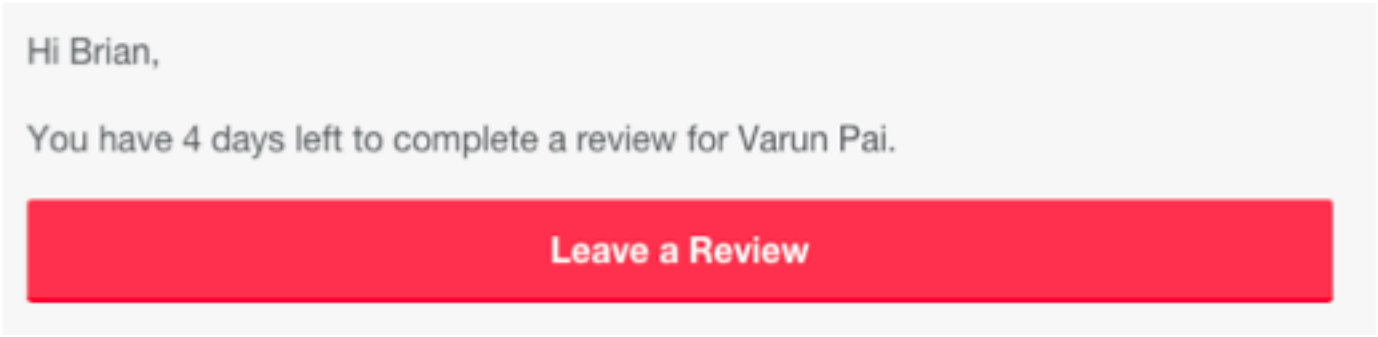
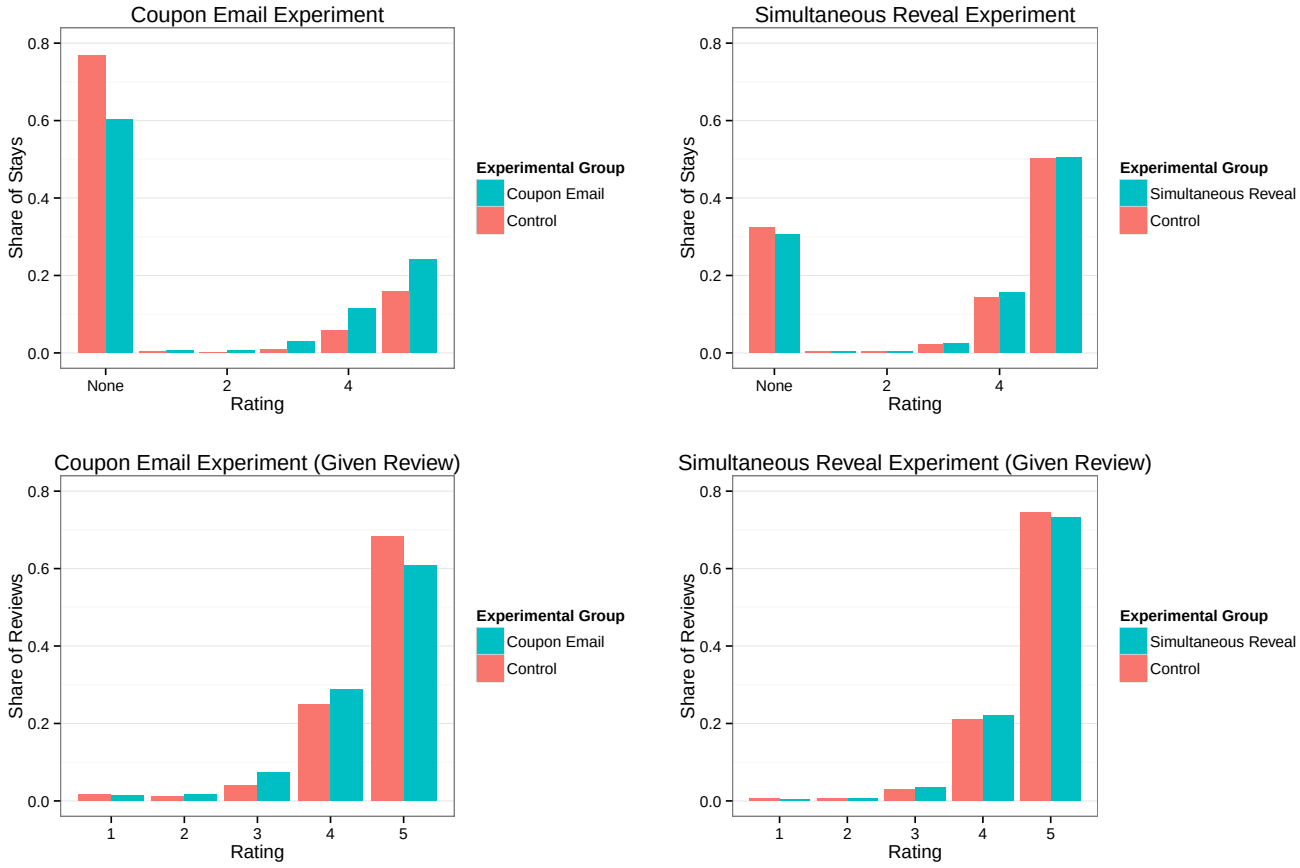
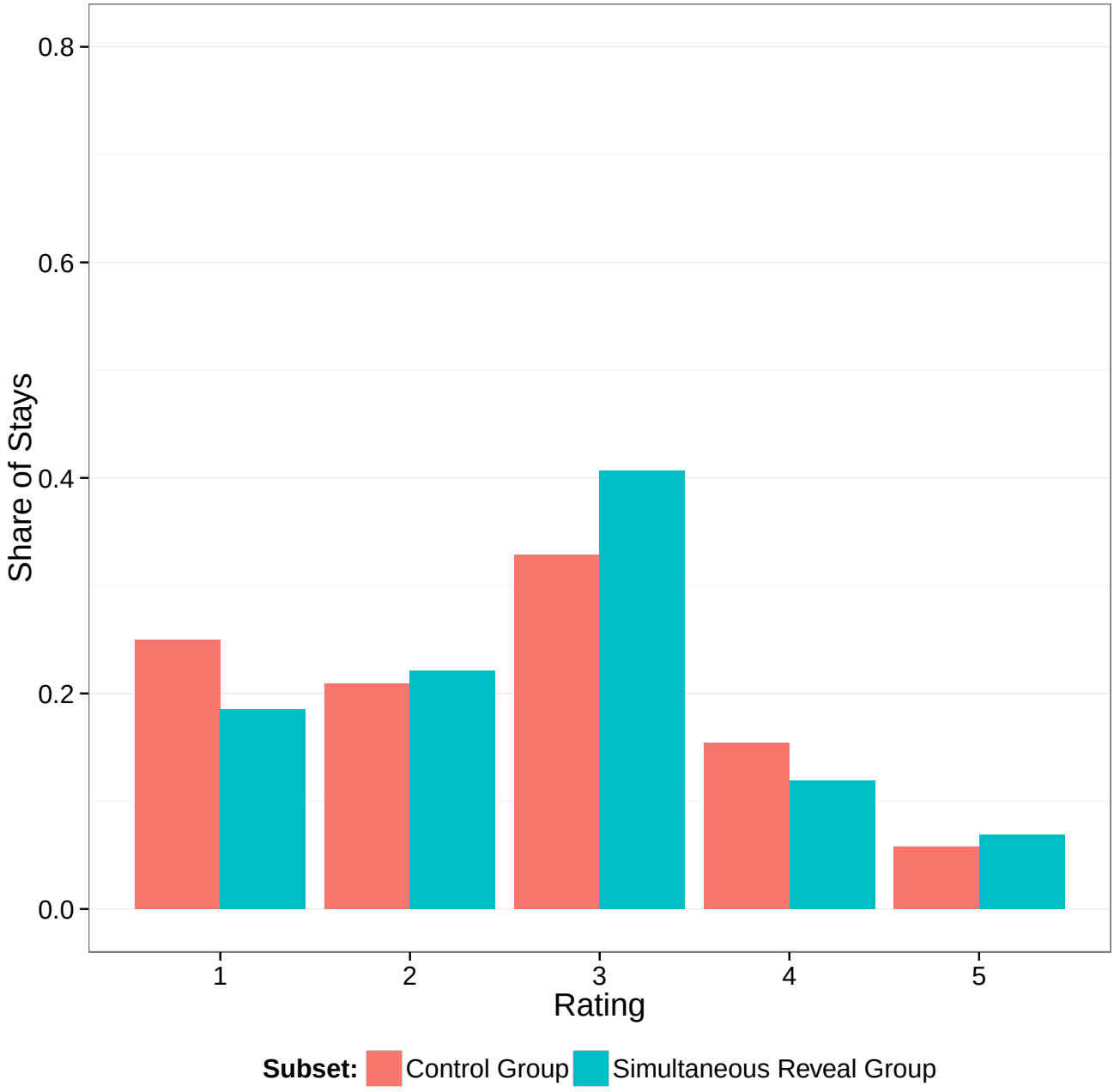


Figure 11: Distribution of Ratings - Experiments



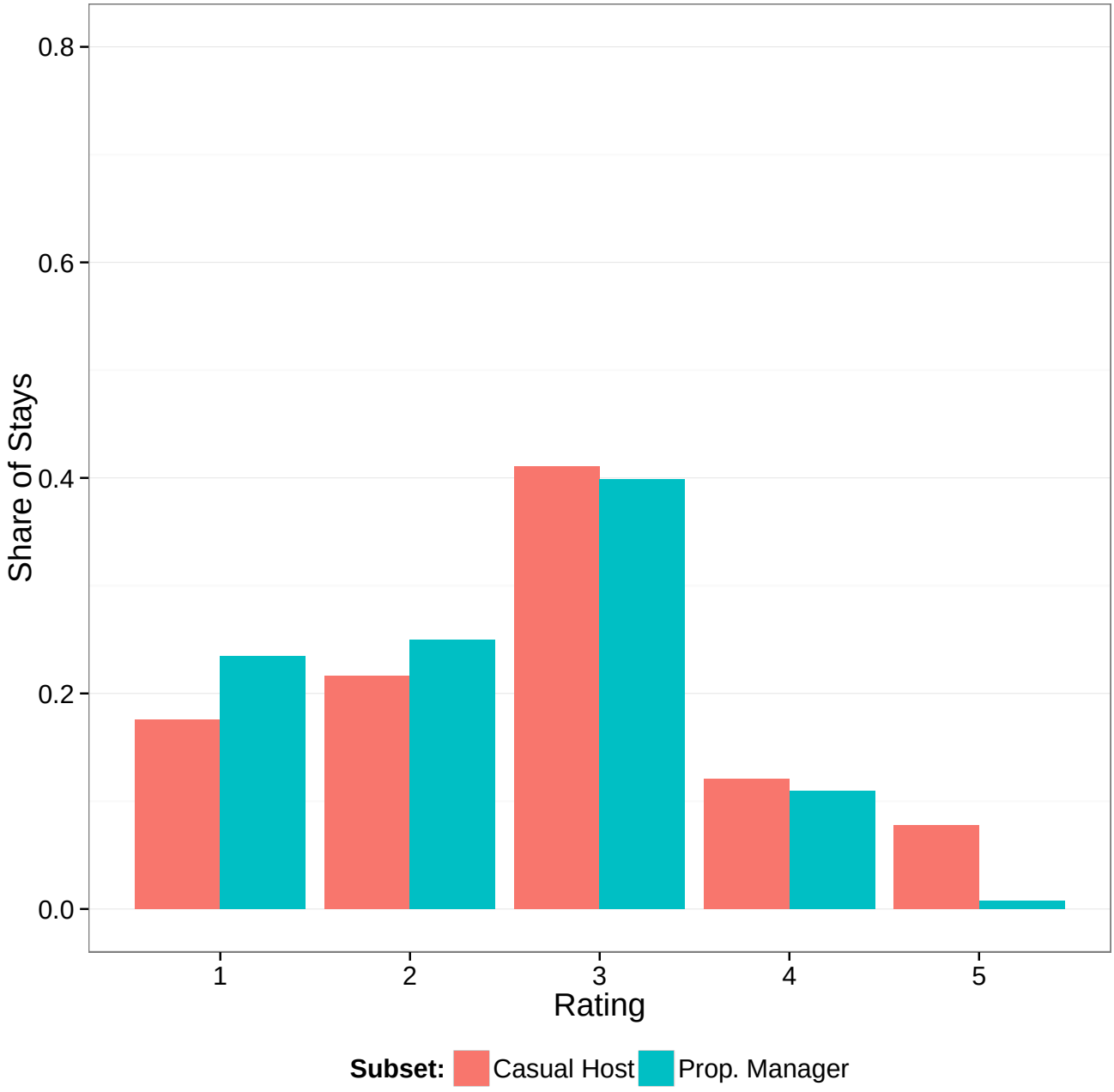
The above figure displays the distribution of ratings in the control and treatment groups for the Simultaneous Reveal Experiment and for the Incentivized Review Experiment. Row 1 displays the distribution of reviews while row 2 displays the distribution of ratings conditional on a review.

Figure 12: Ratings When Guest Does Not Recommend



The above figure displays the distribution of submitted ratings in the control and treatment groups of the simultaneous reveal experiment. Only reviews for which the guest anonymously does not recommend the host are included. Only the first stays for a given host in the experimental time frame is included.

Figure 13: Ratings When Guest Does Not Recommend - Simultaneous Reveal



The above figure displays the distribution of submitted ratings in the treatment groups of the simultaneous reveal experiment. Only reviews for which the guest anonymously does not recommend the host are included. Only the first stays for a given host in the experimental time frame is included.

10 Tables

Table 1: Summary Statistics: Simultaneous Reveal Experiment

	Control		Treatment	
	Guest	Host	Guest	Host
Reviews	0.649	0.716	0.673	0.786
Five Star	0.742	NA	0.726	NA
Recommends	0.974	0.989	0.970	0.990
High Likelihood to Recommend Airbnb	0.766	NA	0.759	NA
Overall Rating	4.675	NA	4.660	NA
All Sub-Ratings Five Star	0.500	0.855	0.485	0.840
Private Feedback	190.749	0.331	188.839	0.330
Feedback to Airbnb	0.125	0.078	0.132	0.085
Median Review Length (Characters)	330	147	336	148
Negative Sentiment	0.161	NA	0.181	NA
Median Private Feedback Length (Characters)	131	101	130	88
First Reviewer	0.337	0.499	0.325	0.527
Time to Review (Days)	3.323	2.689	2.957	2.458
Time Between Reviews (Hours)	64.857	NA	47.906	NA
Num. Obs.	59981	59981	60603	60603

The averages are taken for a sample of trips between 5-11-2014 and 6-11-2014. They do not necessarily represent the historical and current rates of reviews on the site, which differ over time due to seasonality and changes to Airbnb policy. "All Sub-Ratings Five Star" is an indicator variable for whether cleanliness, communication, accuracy, location, value, check-in, and house rules ratings are all 5 stars. "First Reviewer" is an indicator variable for whether the individual submitted the first review for the trip.

Table 2: Determinants of Guest Reviews

	Reviewed	
Avg. Review Rating	0.068***	0.067***
	(0.006)	(0.006)
No Reviews	0.354***	0.351***
	(0.029)	(0.030)
Num. Reviews	0.011***	0.011***
	(0.001)	(0.001)
Num. Trips	−0.008***	−0.008***
	(0.0004)	(0.0004)
Customer Service	−0.125***	−0.123***
	(0.022)	(0.022)
Private Room	−0.003	−0.005
	(0.005)	(0.005)
Shared Room	−0.063***	−0.057***
	(0.017)	(0.017)
New Guest (Organic)	0.044***	0.043***
	(0.008)	(0.008)
Exp. Guest (Marketing)	0.093***	0.094***
	(0.011)	(0.011)
Exp. Guest (Organic)	0.106***	0.106***
	(0.008)	(0.008)
Num. Guests	−0.007***	−0.008***
	(0.001)	(0.001)
Nights	−0.001***	−0.001***
	(0.0003)	(0.0003)
US Guest	−0.0004	−0.004
	(0.004)	(0.005)
Checkout Date	0.001***	0.001***
	(0.0002)	(0.0002)
Price per Night	−0.017***	−0.019***
	(0.003)	(0.003)
Constant	−8.732**	
	(3.492)	
Market FE:	No	Yes
Observations	59,788	59,788
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 3: Experimental Validity Check

Variable	Experiment	Difference	Mean Treatment	Mean Control	P-Value	Stars
Experienced Guest	Incentivized Review	-0.007	0.757	0.764	0.122	
US Guest	Incentivized Review	-0.0002	0.233	0.233	0.956	
Guest Tenure (Days)	Incentivized Review	1.706	226.025	224.319	0.587	
Host Experienced	Incentivized Review	-0.001	0.349	0.350	0.771	
US Host	Incentivized Review	-0.003	0.198	0.201	0.429	
Host is Prop. Manager	Incentivized Review	0.002	0.269	0.267	0.657	
Entire Property	Incentivized Review	0.005	0.696	0.691	0.285	
Host Reviews Within 9 Days	Incentivized Review	0.007	0.491	0.485	0.203	
Observations	Incentivized Review	0.002			0.498	
Experienced Guest	Simultaneous Reveal	0.002	0.704	0.702	0.392	
US Guest	Simultaneous Reveal	-0.001	0.286	0.287	0.735	
Guest Tenure (Days)	Simultaneous Reveal	-2.323	267.814	270.138	0.214	
Host Experienced	Simultaneous Reveal	-0.002	0.811	0.813	0.313	
US Host	Simultaneous Reveal	0.001	0.264	0.263	0.601	
Host is Prop. Manager	Simultaneous Reveal	0.001	0.082	0.081	0.365	
Entire Property	Simultaneous Reveal	-0.0003	0.671	0.672	0.899	
Reviewed Listing	Simultaneous Reveal	-0.004	0.762	0.766	0.103	
Observations	Simultaneous Reveal	0.003			0.073	*

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 4: Summary Statistics: Incentivized Review Experiment

	Control		**Treatment**	
	Guest	Host	Guest	Host
Reviews	0.230	0.596	0.392	0.607
Five Star	0.665	NA	0.589	NA
Recommends	0.818	0.986	0.799	0.987
High Likelihood to Recommend Airbnb	0.725	NA	0.706	NA
Overall Rating	4.565	NA	4.452	NA
All Sub-Ratings Five Star	0.444	0.803	0.378	0.812
Private Feedback	205.350	0.284	200.566	0.292
Feedback to Airbnb	0.123	0.099	0.140	0.097
Median Review Length (Characters)	346	122	300	126
Negative Sentiment	0.215	NA	0.231	NA
Median Private Feedback Length (Characters)	134	93	127	97
First Reviewer	0.069	0.570	0.162	0.548
Time to Review (Days)	16.931	4.888	12.471	4.859
Time Between Reviews (Hours)	279.220	NA	206.573	NA
Num. Obs.	18604	18604	18735	18735

The averages are taken for a sample of trips between 5-11-2014 and 6-11-2014. They do not necessarily represent the historical and current rates of reviews on the site, which differ over time due to seasonality and changes to Airbnb policy. "All Sub-Ratings Five Star" is an indicator variable for whether cleanliness, communication, accuracy, location, value, check-in, and house rules ratings are all 5 stars. "First Reviewer" is an indicator variable for whether the individual submitted the first review for the trip.

Table 5: Effect of Coupon Treatment on Five Star Ratings

	(1)	(2)	(3)	(4)	(5)
Treatment	-0.076^{***}	-0.075^{***}	-0.109^{***}	-0.069^{***}	-0.064^{***}
	(0.009)	(0.009)	(0.026)	(0.009)	(0.016)
Guest Judiciousness			-0.056^{*}		
			(0.032)		
Treatment * Guest Judiciousness			-0.025		
			(0.043)		
Host Rev. First					0.090^{***}
					(0.016)
Treatment * Host Rev. First					0.009
					(0.020)
Guest Characteristics	No	Yes	Yes	Yes	Yes
Listing Characteristics	No	No	No	Yes	Yes
Observations	11,578	11,578	1,439	11,578	11,578

The table displays results of a regression predicting whether a guest submitted a 5 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 property manager, price, room type of the listing, and listing region. *p<0.10, ** p<0.05, *** p<0.01

Table 6: Magnitudes of Experimental Treatment Effects

	Experiment:			
	Incentivized Review	Incentivized Review (Adjusted)	Simultaneous Reveal	Simultaneous Reveal (Non-reviewed Listings)
Dependent Variable:	(1)	(2)	(3)	(4)
Reviewed	0.166***	0.064	0.024***	0.012**
Five Star	-0.085***	-0.021	-0.016***	-0.010
Recommends	0.013	-0.005	-0.003***	-0.004*
Neg. Sentiment	0.047	0.010	0.019***	0.027***

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 7: Retaliation and Induced Reciprocity - Guest

	Does Not Recommend	Overall Rating < 5	Negative Sentiment
	(1)	(2)	(3)
Treatment	0.002	0.029***	0.032***
	(0.002)	(0.006)	(0.005)
Host Negative Sentiment	0.671***	0.690***	0.383**
	(0.128)	(0.122)	(0.159)
Host Does Not Recommend	0.134	0.055	0.254*
	(0.094)	(0.109)	(0.132)
Treatment * Host Negative Sentiment	−0.631***	−0.719***	−0.454**
	(0.159)	(0.177)	(0.208)
Treatment * Host Does Not Recommend	−0.012	0.279*	0.003
	(0.122)	(0.153)	(0.173)
Guest, Trip, and Listing Characteristics	Yes	Yes	Yes
Observations	18,207	18,207	18,207

The above regressions are estimated for the sample where the host 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 8: Retaliation and Induced Reciprocity - Host

	Does Not Recommend	Negative Sentiment	Does Not Recommend	Negative Sentiment
	(1)	(2)	(3)	(4)
Treatment	−0.001	0.003	−0.001	0.005
	(0.001)	(0.008)	(0.001)	(0.009)
Guest Low Rating	0.266***	0.320***	0.120***	0.169***
	(0.030)	(0.049)	(0.034)	(0.057)
Guest Review Negative Words			0.295***	0.311***
			(0.101)	(0.119)
Guest Does Not Recommend			0.107	0.049
			(0.074)	(0.085)
Treatment * Low Rating	−0.195***	−0.244***	−0.053	−0.102
	(0.034)	(0.060)	(0.041)	(0.077)
Treatment * Review Negative Words			−0.187*	−0.275*
			(0.107)	(0.152)
Treatment * Does Not Recommend			−0.154**	−0.056
			(0.075)	(0.116)
Guest, Trip, and Listing Characteristics	Yes	Yes	Yes	Yes
Observations	13,696	7,821	10,431	7,333

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 9: Fear of Retaliation - Host

	Reviews First	Does Not Recommend (First)	Neg. Sentiment (First)	
	(1)	(2)	(3)	(4)
Treatment	0.028***	0.001*	0.002*	−0.001
	(0.003)	(0.001)	(0.001)	(0.001)
Does Not Recommend				0.616***
				(0.044)
Treatment * Does Not Recommend				0.121**
				(0.055)
Guest, Trip, and Listing Characteristics	Yes	Yes	Yes	Yes
Observations	120,230	61,720	31,975	31,975

The regressions in columns (2) - (4) are estimated only for cases when the host 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 10: Fear of Retaliation - Guest

	Reviews First	< 5 Rating (First)	Neg. Sentiment (First)	Neg. Sentiment (First)
	(1)	(2)	(3)	(4)
Treatment	0.0003	0.004	0.008*	0.011**
	(0.002)	(0.004)	(0.004)	(0.005)
< 5 Rating				0.158***
				(0.009)
Not Recommend		0.679***		0.458***
		(0.006)		(0.022)
Treatment * < 5 Rating				−0.011
				(0.012)
Treatment * Not Recommend				−0.027
				(0.031)
Guest, Trip, and Listing Characteristics	Yes	Yes	Yes	Yes
Observations	37,297	37,295	30,908	28,957

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 11: Determinants of Private Feedback Increase

	Guest Left Private Suggestion for Host		
	(1)	(2)	(3)
Treatment	0.063***	0.036***	0.041***
	(0.003)	(0.004)	(0.007)
Customer Support	0.011	0.030*	0.018
	(0.017)	(0.017)	(0.017)
Guest Recommends		0.076***	0.084***
		(0.003)	(0.003)
Five Star Review			−0.075***
			(0.005)
Recommends * Treatment		0.032***	0.034***
		(0.004)	(0.004)
Five Star * Treatment			−0.010
			(0.007)
Guest, Trip, and Listing Characteristics	Yes	Yes	Yes
Observations	79,476	79,476	79,476

"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, 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 12: Socially Induced Reciprocity - Star Rating

	Overall Rating (1 - 5 Stars)		
	(1)	(2)	(3)
Private Room	0.025***	0.025***	0.020***
	(0.007)	(0.007)	(0.007)
Prop. Manager	−0.085***	−0.082***	−0.093***
	(0.012)	(0.013)	(0.013)
Guest Does Not Rec.	−1.960***	−2.013***	−1.997***
	(0.046)	(0.034)	(0.035)
Private Room * Does Not Rec.	−0.063		
	(0.076)		
Prop. Manager * Does Not Rec.	−0.185*		
	(0.095)		
Low LTR		−0.289***	
		(0.010)	
Private Room * Low LTR		0.025	
		(0.017)	
Prop. Manager * Low LTR		−0.016	
		(0.030)	
Comment to Airbnb			−0.055***
			(0.012)
Private Room * Comment to Airbnb			0.033*
			(0.020)
Prop. Manager * Comment to Airbnb			−0.002
			(0.037)
Guest, Trip, and Listing Characteristics	Yes	Yes	Yes
Market FE	Yes	Yes	Yes
Observations	37,965	33,936	37,965

The outcome in the above regression is the guest's star rating. The sample used is the set of first trip in the treatment group of the simultaneous reveal experiment."Rec." refers to the anonymous recommendation that the guest can submit. "Low LTR" occurs when a guest responds to the likelihood to recommend prompt with a lower than 9 out of 10. "Comment to Airbnb" is an indicator variable for whether the guest submits private feedback to Airbnb. 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 13: Socially Induced Reciprocity - Star Rating (Guest Fixed Effects)

	Overall Rating (1 - 5 Stars)	
	(1)	(2)
Entire Property	−0.005***	−0.006***
	(0.002)	(0.002)
Num. Listing Trips	0.0001**	0.0001**
	(0.00002)	(0.00002)
Property Manager	−0.113***	−0.112***
	(0.002)	(0.002)
Comment to Airbnb	−0.099***	−0.091***
	(0.003)	(0.003)
Effective Review Rate	0.099***	0.097***
	(0.001)	(0.001)
Customer Service	−0.161***	−0.155***
	(0.005)	(0.005)
No LTR		0.807***
		(0.012)
Likelihood to Recommend		0.086***
		(0.001)
Recommends Listing	1.943***	1.892***
	(0.005)	(0.005)
Prop. Mgr. * Num. Listing Trips	0.0002***	0.0002***
	(0.00004)	(0.00004)
Trip Characteristics	Yes	Yes
Market FE	Yes	Yes
Guest FE	Yes	Yes
Observations	2,073,553	2,073,553

The outcome in the above regression is the guest's star rating. The sample used is all trips between April 2014 and November 2014. "No LTR" occurs when the guest does not submit a likelihood to recommend answer. "Comment to Airbnb" is an indicator variable for whether the guest submits private feedback to Airbnb. Additional controls which are not shown include the number of nights and guests for the trip, and whether the guest submitted a recommendation. *p<0.10, ** p<0.05, *** p<0.01

Table 14: Socially Induced Reciprocity - Negative Sentiment

	Does Not Recommend	Negative Sentiment		
	(1)	(2)	(3)	(4)
Private Room	0.0002	−0.015***	−0.016***	−0.016***
	(0.002)	(0.005)	(0.005)	(0.005)
Prop. Manager	0.021***	0.022**	0.020**	0.028***
	(0.004)	(0.010)	(0.010)	(0.009)
Guest Does Not Rec.		0.185***	0.149***	0.098***
		(0.024)	(0.025)	(0.022)
Overall Rating 2		−0.020	−0.0002	0.041
		(0.042)	(0.042)	(0.041)
Overall Rating 3		−0.169***	−0.122***	−0.026
		(0.038)	(0.041)	(0.039)
Overall Rating 4		−0.395***	−0.298***	−0.181***
		(0.039)	(0.044)	(0.042)
Overall Rating 5		−0.508***	−0.374***	−0.281***
		(0.039)	(0.044)	(0.042)
Low Subrating 1			0.210***	0.131***
			(0.039)	(0.035)
Low Subrating 2			0.163***	0.103***
			(0.034)	(0.031)
Low Subrating 3			0.096***	0.042
			(0.028)	(0.026)
Low Subrating 4			0.037	−0.013
			(0.027)	(0.025)
Low Subrating 5			0.007	−0.055**
			(0.027)	(0.025)
Guest, Trip, and Listing Characteristics	Yes	Yes	Yes	Yes
Market FE	Yes	Yes	Yes	Yes
Review Length Polynomial	No	No	No	Yes
Observations	37,965	29,645	29,645	29,645

"Rec." refers to the anonymous recommendation that the guest can submit. "Low LTR" occurs when responds to the likelihood to recommend prompt with a lower than 9 out of 10. "Comment to Airbnb" is an indicator variable for whether the guest submits private feedback to Airbnb. 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 15: 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
All Biases	1.66	0.76	68.44
No Strategic Bias	1.43	0.81	63.00
No Social or Strategic Bias	0.86	0.38	52.69
No Social, Strategic or Sorting Bias	0.45	0.45	41.35
Everyone Reviews	0.45	0.45	12.90

The above table displays three measures of bias under four 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. "All Biases" is the scenario corresponding to the control group of the simultaneous treatment experiment. "No Strategic Bias" corresponds to the treatment group of the simultaneous reveal experiment. "No Social or Strategic Bias" adjusts all mis-reporting rates to be the same as they are for property managers with entire properties. "No Social, Strategic or Sorting Bias" sets the rates of reviews for those with positive and negative experiences to be equivalent (while keeping the overall review rate constant). "Everyone Reviews" displays the measures of bias when the three biases are removed and every guest submits a review.