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

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.

November 13, 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. Therefore, the distribution of submitted evaluations may differ from the distribution of experiences had by market participants. We study the determinants and size of bias in online reviews by using field experiments on Airbnb. We show that reviews are indeed biased. In the first experiment, we induce more consumers to leave reviews by offering them a coupon. We find that the rate of positive reviews falls by 4.3 percentage points in the treatment group. In our second experiment, we remove the possibility of retaliation in reviews by changing the rules of the review system. We show that bias due to strategic reasons is relatively small but that fear of retaliation, retaliation against negative reviews, and reciprocity of positive reviews all cause misreporting. 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 causes the largest bias of all the causes that we consider and represents a major challenge for online marketplaces that intermediate transactions involving social interaction.

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

1 Introduction

Online reviews and reputation scores are increasingly used by consumers to decide what goods to buy and whom to trade with. These reputation systems are especially important for online marketplaces, where economic agents are often anonymous, trade infrequently, and provide heterogeneous services. In such settings, inaccurate reputations can decrease market efficiency by generating worse matches, reducing trust in the system, and reducing the incentive of market participants to exert effort. However, because consumers 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 (e.g. Avery, Resnick and Zeckhauser [1999], Miller, Resnick and Zeckhauser [2005]). What then explains why people leave reviews and whether those reviews are accurate? We use two field experiments and proprietary data from Airbnb, a large online marketplace for accommodations, to study these questions.

Our first experiment offers a subset of buyers a coupon to submit a review. Our second experiment, removes the possibility of strategic reviewing by changing the review system to simultaneously reveal reviews rather than to reveal each review immediately after submission. Both of these experiments make the reputation system less biased by increasing review rates and by lowering positive review rates. We use the results of these experiments to test and quantify the relative importance of theories of reviewing behavior. We show that reviewers are influenced by the effort of reviewing, behavioral motivations, and strategic considerations.

With regard to buyer reviews of sellers, we find that the largest source of bias in reviews is due to *socially induced reciprocity*, which occurs when people socially interact with each other while transacting and consequently omit negative public feedback. Furthermore, we find that the experiences of those who review are not representative of the experiences of everyone who transacts. We show that buyers with mediocre experiences are less likely to review than those with positive experiences. However, contrary to some theories proposed in the literature,[1] sorting into reviewing is not primarily driven by the fear of retaliation but

by a general dislike of writing negative reviews.

We do find evidence of strategic behavior, where the fear of retaliation and the possibility of induced reciprocity lead to misreporting of experiences by buyers and sellers. Furthermore, we show that retaliation and induced reciprocity do occur, making these strategic actions warranted. However, this behavior is a relatively minor cause of bias for buyer reviews of sellers. For seller reviews of buyers, the fear of retaliation decreases the rate of negative sentiment in reviews by 12 percentage points conditional on the seller having a bad experience.

The setting of this paper is Airbnb, a prominent marketplace for accommodations where guests (buyers) stay in the properties of hosts (sellers) and review each other afterwards. Reputation is particularly important for transactions on Airbnb because guests and hosts interact in person, often in the primary home of the host. Guests must trust that hosts have accurately represented their property on the website, while hosts must trust that guests will be clean, rule abiding, and respectful. Airbnb's reputation system displays two empirical regularities that are seen in many other online reputation systems (e.g. Dellarocas and Wood [2007], Nosko and Tadelis [2014]): many participants do not leave a review, and most reviews are positive. Over 30% of guests in our sample do not leave a review and over 70% of reviews by guests are five stars, the highest rating possible.

We first show that the average ratings on Airbnb are informative about the experiences of guests and hosts. Guests and hosts who call customer support and provide private feedback to Airbnb, leave lower ratings. Furthermore, anonymous feedback is correlated with public feedback ratings. However, ratings and review text are not fully informative. For example, 50% of guests who did not recommend their host in an anonymous prompt submitted a five star rating.

[2]For example, Dellarocas and Wood [2007] claim: "it is widely believed (though, so far, not rigorously proven) that many traders choose to remain silent because they are afraid that, if they report their negative experience, their partner will "retaliate" by posting negative feedback for them as well."

Our first experiment is designed to test whether guests who review have different experiences than guest who don't. We offer guests a $25 coupon to leave a review to an non-reviewed listing. We find that the rate of reviews in the treatment group is 6.7 percentage points higher and that the share of those reviews that are five star is 4.3 percentage points lower. This is the first experimental evidence that reviewers with worse experiences are less likely to leave reviews than those with positive experiences. Furthermore, the increase in negative reviews comes from 3 and 4 star ratings rather than the 1 star ratings corresponding to very negative experiences.

Our second experiment removes the possibility of strategic reviewing responses by hiding any feedback until both the buyer and seller have left a review (or the review time 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 by 2.4 percentage points while decrease the share of five stars reviews by 1.6 percentage points. On the host side, the simultaneous reveal treatment increases review rates by 7 percentage points.

We show that the overall treatment effect masks strategic responses by guests and hosts. We first show that guests respond to the content of initial host reviews. Guests in the treatment group leave higher ratings than those in the control when the host leaves a negative review first, and lower ratings when the host leaves a positive review first. This demonstrates two effects. First, guests are induced to leave positive reviews by positive host reviews. When the guests no longer see the review content, they are 3.2 percentage points more likely to leave negative text in their reviews. Second, guests retaliate against negative host reviews. Amongst guests who receive negative host reviews, those in the control are over 30 percentage points more likely to respond with negative review text than those in the treatment. We use the same empirical strategy for host reviews of guests when guests review first. We find that, while hosts do retaliate, they are not induced to reciprocate by positive guest reviews.

We then test whether first reviewers understand the effect of their reviews on subsequent

reviews. We find that the treatment induces both guests and hosts to leave more negative review text in a first review. Furthermore, hosts who do not recommend the guest in an anonymous review prompt are 12 percentage points more likely to express that attitude in their review text when in the treatment. Guests also are more likely to leave a negative first review on the treatment, however, this effect is not driven by guests who do not recommend the host. Instead, the fear of retaliation has the largest effect on the behavior of guests who have a good, but not perfect experience.

Lastly, we test whether social interaction during a transaction affects bias in reviews. Our empirical strategy relies on the fact that different types of trips on Airbnb entail different levels of social interaction. For example, a trip to a private room within a larger property is more likely to entail a social interaction than a trip in which an entire property is rented. Guests and hosts might interact with each other in this setting while using the living room or kitchen, or when walking to their rooms. On the other hand, trips to entire properties do not involve such interactions. Similarly, trips to properties with professional managers are less likely to result in social interactions because property managers often manage listings remotely.

We cannot simply compare ratings between these types of trips, because trips differ in a variety of ways unrelated to social interactions. Instead, our identification of socially induced reciprocity uses the difference between public and private review ratings. We find that there is have a higher chance of mismatch between public review text and private recommendations in reviews of trips with a higher likelihood of social interactions. Conditional on the guest not recommendation a listing, reviews of private rooms are 11 percentage points less likely to involve negative sentiment than reviews of entire properties managed by property managers. Furthermore, our empirical strategy likely understates the degree of socially induced reciprocity because almost all trips involve some amount of communication. Socially induced reciprocity is likely to be important for evaluations in setting other than Airbnb, such as labor markets and other peer-to-peer marketplaces. For example, employers often use

written recommendations and back-channel references from an applicant's prior co-workers when deciding to make an offer.[2] If reference givers have socialized with the candidate, they may omit negative information and overstate positive accomplishments in their evaluations. Social conversation is also common during transactions in other peer-to-peer marketplaces such as Uber (cabs), Taskrabbit (errands), and Odesk (labor).

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 [1995], Andreoni and Rao [2011]). Anonymity is another important factor in giving behavior. For example, Hoffman et al. [1994], Hoffman, McCabe and Smith [1996] find that giving decreases with more anonymity and increases with language suggesting sharing. Our results show that these laboratory results carry over to reviewing behavior. We find that positive review rates decrease with greater social distance and with increased anonymity of reviews. Another related literature shows that participation in giving games is actually an endogenous variable (e.g. Malmendier, te Velde and Weber [2014], Lazear, Malmendier and Weber [2012], and DellaVigna, List and Malmendier [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 reviewee.

Our paper complements the growing literature on the effects of reviews and the design of reputation systems. The literature on reviews has shown that reviews and review ratings causally affect demand across a variety of settings.[3] Furthermore, Fradkin [2014] shows

[2]In an online labor market setting, LinkedIn allows users to write publicly displayed recommendations and skill endorsements for each other. One author of this paper was endorsed for technical skills such as "Econometrics" by a distant acquaintance who has no technical knowledge.

that guests on Airbnb value higher rated listings more and that hosts reject reviewed guests less. However, given that biased public reviews affect market outcomes, does their bias matter? Horton [2014] provides the most convincing evidence that review bias does matter for market outcomes. He experimentally shows that demand by employers on Odesk changes when anonymized feedback is shown in addition to public feedback. His results demonstrate that market participants are aware of bias in public reviews and react when less biased information is available. Another benefit of less biased feedback is that it reduces moral hazard amongst traders (e.g. Hui et al. [2014]).

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), while Nosko and Tadelis [2014] propose a heuristic proxy for quality, the effective positive percentage (EPP) ratio. Nosko and Tadelis [2014] document that purchases from Ebay sellers with lower EPP rates are more likely to result in bad buyer experiences. They propose alleviating this market failure by manipulating the search ranking algorithm to favor listings with higher EPP. We validate EPP as a useful measure of seller quality by experimentally inducing additional reviews and showing that Airbnb listings with lower EPP receive lower ratings in the experiment.

Our paper also provides a test of proposed reputation system designs in the literature. Bolton, Greiner and Ockenfels [2012] provide observational evidence from Ebay and Rentacoder.com that strategic motivations are important in two-sided review systems. They propose a "blind" review system to remove strategic bias 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, we find that non-strategic sources of bias are even more important in our setting. Our coupon intervention also reduced bias but

[3] Pallais [2014] uses experiments to show that reviews affect demand for workers on Odesk. Luca [2013] shows that Yelp star ratings affect demand using a regression discontinuity design. Cabral and Hortaçsu [2010] use panel data to show that reputation affects exit decisions by firms on Ebay.

coupons are too expensive and prone to manipulation to be used broadly. 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] shows that rebates induce reciprocity in buyers and actually increase the bias in reviews.

In the next section we describe our setting in more detail and provide descriptive statistics about the review system. In section 3 we document sorting bias and show how the incentivized review experiment reduces that bias. In section 4 we describe the simultaneous reveal experiment and its reduced form effects. In section 5, we document strategic incentives in reviewing and in section 6 we document social reciprocity. We discuss the size of review bias in section 7 and then conclude.

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. Furthermore, a reminder is automatically sent by email if a person has not reviewed within 9 days or if the counterparty has left a review.

Airbnb's review process for listings involves 3 pages consisting of public, private, and

anonymous question prompts (shown in Figure 1). Guests are initially prompted to leave textual feedback consisting of a public revealed comment, a 1 to 5 star rating[4], and private comments to their hosts (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 sub-scores are displayed on each listings 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 their trip using a text box and a "likelihood to recommend" (LTR) question prompt.[5]

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 reviewed guests experience lower rejection rates by subsequent hosts, conditional on observable characteristics. 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 overall characteristics of reviews on Airbnb. We use data for 115 thousand trips between May 10, 2014 and June 12, 2014 that are in the control group of the subsequent experiments.[6] The summary statistics for these trips are seen in columns

[2]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.

[3]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".

1 and 2 of Table 1. Turning first to reviews rates, 67% of trips result in a guest review and 72% result in a host review.[7] 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 their Airbnb website more frequently than guests. Second, hosts have more to gain from inducing a positive review by a guest and therefore tend to review first.

We first consider guest reviews of hosts. Figure 6 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. Furthermore, there is no spike in the distribution for 1 star reviews as seen on sites like Amazon.com. Guest reviews on Airbnb are overwhelmingly positive compared to other hotel review websites. The rate of five star reviews is 44% on and 31% on TripAdvisor.[8] To the extent that both Airbnb and hotel sell accommodations services, it is unlikely that 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 a different set 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 based on several sub-categories. 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. However, even if guests leave a lower than 5 star rating, they often recommend a

[4]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.

[5]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.

listing. Eighty-eight percent of guests responded that they would recommend a listing, even though the answer was clearly marked as anonymous in the prompt. This shows that most guests have a good experience on Airbnb.

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 the only part of the review that is permanently associated with the individual reviewer. Previous research has shown that review text contains information about the characteristics of goods that is important for consumer decisions (Archak, Ghose and Ipeirotis [2011]). Figure 8 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 6 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 50% of cases where the guest does not recommend the host, the guest leaves a five star rating. Therefore, guests are lying about the quality of their experiences. One worry is that guests do not understand the review prompt. However, the fact that fewer than 5% of guests recommend a listing when they leave a lower than a four star rating suggests that guests indeed understand the review prompt. Lastly, 20% of recommendations are associated with a four star ("Great") rating

[8]These averages are from Mayzlin, Dover and Chevalier [2014].

by the guest. Therefore, some guests consider a four star rating to be a positive, rather than negative, overall experience. The heterogeneity of interpretations of review prompts is an important consideration in designing reputation systems that we do not consider 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 displayed in Figure 8. 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% for stays in which hosts call customer support. We turn to the hosts' review text to determine the characteristics of a bad guest. Figure 9 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 (e.g. Dellarocas and Wood [2007]). For example, if potential reviewers value sharing positive experiences more than negative experiences, then the observed reviews on the website would be upwardly biased. We document a large sorting bias in the Airbnb review system and show how paying 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 then describe and evaluate an experiment that induces guests to leave reviews.

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 confirms that worse listings and experiences are associated with lower review rates. Listings with lower ratings and lower historical review rates per trip are 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. Lastly, more expensive trips are less likely to be followed by a review.

Guest characteristics also influence the probability that a review is submitted. New

guests and guests who found Airbnb through online marketing efforts are less likely to leave reviews after a trip. The fact that new guests and those acquired through marketing might be due to one of two reasons. First, experienced users who found Airbnb through their friends might be more committed to the Airbnb ecosystem and might be more likely to feel an obligation to review. New users and users acquired through online marketing might have worse experiences on average, either because they picked a bad listing or because they had different expectations about using Airbnb.

3.1 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 listing level). The treatment group received an email offering a $25 Airbnb coupon while the control group received a normal reminder email (shown in Figures 4 and 5).

Since the treatment affected the probability of a review, there were more trips in the control group than in the treatment group. We therefore limit our analysis to only the first trip in the experiment per listing. 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 assignment is valid.

Table 4 displays the review related summary statistics for this experiment in the treat-

ment and control groups. 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 at all. The treatment increases the review rate in this sample by 70% and it decreases the share of five star reviews by 11%. Figure 7 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 two percentage point decrease in the recommendation and likelihood to recommend Airbnb responses.

The effect of this experiment on the review ratings might be due to one of three reasons. The first reason 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 by in Table 6. Column (1) of Table 6 displays the baseline treatment effect of the experiment without any control. Column (2) adds in control for guest origin, experience, and trip characteristics. These 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) limits the sample to experienced guests and adds controls for the historical judiciousness of a guest when leaving reviews and an interaction between the historical judiciousness of the guest and the treatment. The guest judiciousness variable is the negative guest specific fixed effect on a regression of ratings on guest and listing fixed effects.[9] Lower values of guest judiciousness occur when guests always leave high ratings for hosts. 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 additional negative reviews are not due to the fact that different types of guests are induced to review. In column (4), we test whether more negative reviews are driven by listing composition. Adding controls for listing type, location, price, and number of unreviewed stays increases the treatment effect to 6.4 percentage points. Lastly, we test whether the delayed review timing by guests in the experimenet is driven by fear of host retaliation. If fear of retaliation affects the review ratings, then we would expect the ratings to be lower in the treatment when the host has already reviewed the guest. We add controls for whether the host reviewed the guest first in Column (5). Overall, guest ratings are 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 treatment to omit reviews.

4 The Simultaneous Reveal Experiment

Our second experiment changes the timing by 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 had the problem that the second reviewer had the opportunity to either retaliate against or reciprocate the first review. To the extent that retaliation or reciprocation reflect strategic concerns rather than the quality of the experience, this mechanism was biasing the review system. We designed an experiment with the goal of reducing these strategic concerns and making the review system more accurate.

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 5% holdout group. The first

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

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. We modified the maximum time to review because we did not want some reviews to be hidden for an entire month. For the rest of this paper we use the "short expiration" treatment as the control and the "simultaneous reveal" treatment as the experimental as the treatment.[10]

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 percentage points, while the recommendation rate increases by 1 percentage point. Furthermore, the anonymous recommendation responses by hosts stay at 99% of all reviews. However, the text of the submitted reviews does change. There rate of negative sentiment 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 6pp increase for guests and a 2pp increase for hosts. Lastly, hosts in the experiment review quicker and are 3pp more likely to review first in the treatment group. In the next two sections we propose a theory of reviewing behavior and test it using our experiments.

5 Strategic Motivations of Reviewers

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 the actions of the other reviewer or in response to the other review. Intrinsic motivations influence a reviewer to leave a particular type of review independently of the potential response or prior reviewing actions of the other reviewer.

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 because they are either afraid of retaliation or they want to induce the second reviewer to leave a positive review. 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.

The simultaneous reveal experiment allows us to test the importance of strategic motivations for reviewers. 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:

$$y_{gl} = \alpha_0 t_l + \alpha_1 FRN_{gl} + \alpha_2 FN_{gl} + \alpha_3 t_l * FRN_{gl} + \alpha_4 t_l * FN_{gl} + \beta' X_{gl} + \epsilon_{gl} \quad (1)$$

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

If there is induced reciprocity the we would expect α_0 to be positive, because guests in the treatment group do not see the content of the first review. Second, if there is retaliation against negative host reviews, we would expect α_2 to be positive and α_4 to be negative. That is, guests retaliate against negative host review text only when they can see it. 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 little effect of the treatment on this relationship.

Table 7 displays estimates of Equation 1 for cases when the guest reviews second. Column

(1) shows the estimates when the outcome variable is whether the guest does not anonymously 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 9 appears in the review text. The coefficient on host negative sentiment is .68 and the interaction with the treatment is -.72. 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 88% baseline rate of recommendations. Columns (2) and (4) 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 on retaliation using these outcome variables.[11] 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. Columns (3) and (5) limit the estimation sample to only those cases when the guest did not recommend the host. The coefficients on host sentiment in column (3) are small and insignificant, demonstrating that guests who do not recommend are honest in their star ratings. However, column (5) shows that those guests do omit negative sentiment from review text.

Next, we consider the same specification for cases when hosts review second. Figure 8 displays estimates for two outcomes: whether the host does not recommend and whether the host uses negative sentiment. For all specifications, 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 by a guest makes hosts 3.8 percentage points

[11]The size of the retaliatory response is smaller for negative sentiment, but this is due to measurement error in the classification of guest reviews.

less likely to recommend and 4.3 percentage points more likely to leave negative review text (defined across all host reviews regardless of the host's recommendation). In specifications (3) and (4), we look at three types of initial guest feedback: recommendations, ratings, and negative sentiment conditinoal 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 20 percentage points, while the coefficients on guests' ratings' are an order of magnitude smaller and not always statistically significant.

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:

$$y_{gl} = \alpha_0 t_l + \alpha_1 DNR_{gl} + \alpha_2 DNR_{gl} * t_l + \epsilon_{gl} \tag{2}$$

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 counterparty. 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 the results of Equation 2 for first reviews by hosts. Column (1) displays the effect of the treatment on the probability that a host reviews first. Hosts are 2.9 percentage points more likely to review in the treatment. This demonstrates that hosts change their timing of reviews to a greater extent than guests. This effect is expected since hosts have more reason to be strategic than guests. While hosts rely on Airbnb reviews for income,

guests can substitute to a hotel if they receive a bad review. Column (2) studies 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. 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 a 1.5 percentage points increases in the overall rate of negative text in first host reviews. Furthermore, there is an additional 12 percentage points increase in the rate of negative text if the host does 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.

We run the same set of specifications for guests' first reviews in Table 10. Column (1) confirms that guests are less likely to review first in the treatment. Column (2) shows that there is no difference in whether guests recommend in the treatment and control. Columns (3) and (4) display the effects of the treatment on the likelihood that guests leave negative sentiment in their reviews of hosts. There is an overall increase in sentiment between .7 and 1.1 percentage points, however this rate is not affected by whether the guest recommends the listing. We interpret this result as follows. Guests with bad experiences are not more afraid of retaliation than other guests. Furthermore, guests with positive, but not perfect, experiences omit negative text because they do not think that the risk of retaliation is worth the benefit of being honest.

We confirm this theory by studying the effect of the treatment on private feedback. Guests have the ability to leave suggestions for a host to improve the listings. However, if guests are afraid of retaliation, then they may choose not to leave this private feedback. 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 6.2pp, 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 entire effect of the treatment on suggestions comes from

guests who recommend the host. Therefore, guests with good but not perfect experiences are influenced by the fear of retaliation.

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. Experimental results show that social communication can affect reviewing behavior for a variety of reasons including empathy (e.g. Andreoni and Rao [2011]), social pressure (e.g. Malmendier, te Velde and Weber [2014]), and the increased potential for repeated interactions.

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 professional property manager. 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.

We cannot simply compare ratings for these types of listings because these listings may differ in other ways that affect reviews. Instead, our identification strategy 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 we expect guests to leave higher public ratings conditional on the anonymous ratings they submit. Our specification to test for this effect is:

$$y_{gl} = \alpha_0 PR_l + \alpha_1 PM_l + \alpha_2 DNR_{gl} + \alpha_3 PR_l * DNR_{gl} + \alpha_4 PM_l * DNR_{gl} + \beta' X_{gl} + \epsilon_{gl} \quad (3)$$

where y_{gl} is a negative review outcome, PR_l and PM_l are indicators for whether listing l is a private room and management by a property manager, DNR_{gl} is an indicator for whether the did not recommend the listing, and X_{gl} are guest and trip characteristics. If socially induced reciprocity occurs then we expect α_3 to be negative because guests to private rooms should leave less negative feedback. Furthermore, we expect α_4 to be positive because property managers induce less reciprocity in guests.

Tables 12 and 13 display the results of the above specification for negative sentiment and overall ratings which are less than 5 stars. Column (1) in both specifications dispays teh the results of the baseline specification. Overall, guests who do not recommend are 16 percentage points more likely to leave negative feedback and 26 percentage points less likely to leave a 5 star rating. However, for guests to private rooms, this effect decreases by 4 percentage points and 3.6 percentage points respectively. The effect for property managers is also in the expected direction. When the guest does not recommend, reviews of property managers are 3.9pp more likely to include negative sentiment and 4.7 percentage points more likely to have a low rating. Columns (2) and (3) of both tables repeat this test with other measures of guest dissatisfaction: whether a guest states a low likelihood to recommend Airbnb and whether the guest leaves private feedback for Airbnb. The coefficients on the interaction between negative experience and room type for these specifications are also of the predicted

sign. We therefore conclude that socially induced reciprocity does affect reviewing behavior.

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 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 of bias 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 uniformative 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 of bias 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 (e.g. 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. Furthemore, suppose that there are only two types of reviews, positive (s_g) and negative. Then the share of submitted ratings that are positive is:

$$\bar{s} = \frac{gPr(r|g)Pr(s_g|g,r) + (1-g)Pr(r|n)Pr(s_g|n,r)}{Pr(r)} \tag{4}$$

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

$$B_{avg} = (1-g)\frac{Pr(r|n)Pr(s_g|n,r)}{Pr(r)} - g(1 - \frac{Pr(r|g)Pr(s_g|g,r)}{Pr(r)}) \tag{5}$$

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).[12] 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, $Pr(r|g) \neq Pr(r|b)$. Therefore, we cannot simply use the rates of recommendations in the data to back out g. Instead, we calibrate g is by using the recommendation rates from the incentivized review experiment, which eliminates some of the 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 reweight the baseline rate of recommendation for listings with 0 reviews

by the relative rates of recommendations in the overall sample.

$$\hat{g} = s_{0,ir,sr} \frac{s_{all,sr}}{s_{0,c,sr}} \tag{6}$$

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. To do so, we use the empirical rates of mis-reporting in each of the scenarios described in the next section.

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

$$B_{mis} = \frac{N_{p|n} + N_{n|p}}{N_{rev}} \tag{7}$$

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}:

$$B_{neg} = 1 - \frac{N_{n|n}}{N_{all}(1-g)} \tag{8}$$

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.

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

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 social bias in the reviewing process. To do so, we let $\widehat{Pr(s_g|n,r)}$ 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 14 displays each measure of bias for all 4 scenarios described above. We first turn

to the case when all biases operate (row 1). In this scenario, positive reviews occur 7.8% more of the time than positive experiences. Furthermore, 6% of all reviews mis-represent the quality of a guests experience and 84% of negative experiences are not reported in textual reviews. Removing strategic considerations changes these numbers by less than .005 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 6.3%, while the share of misreported reviews is 4.7% of all reviews. This represents a drop in bias that is an order of magintude 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.

Lastly, in row 4, we remove sorting bias. The average bias falls an additional .6 percentage points and the share of negative experiences missing drops to 69% 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. Furthermore, because a large of reviews are non-recommendations in this scenario, the share of mis-reported reviews actually increases when sorting is removed.

8 Conclusion

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, 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." *Manage. Sci.*, 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

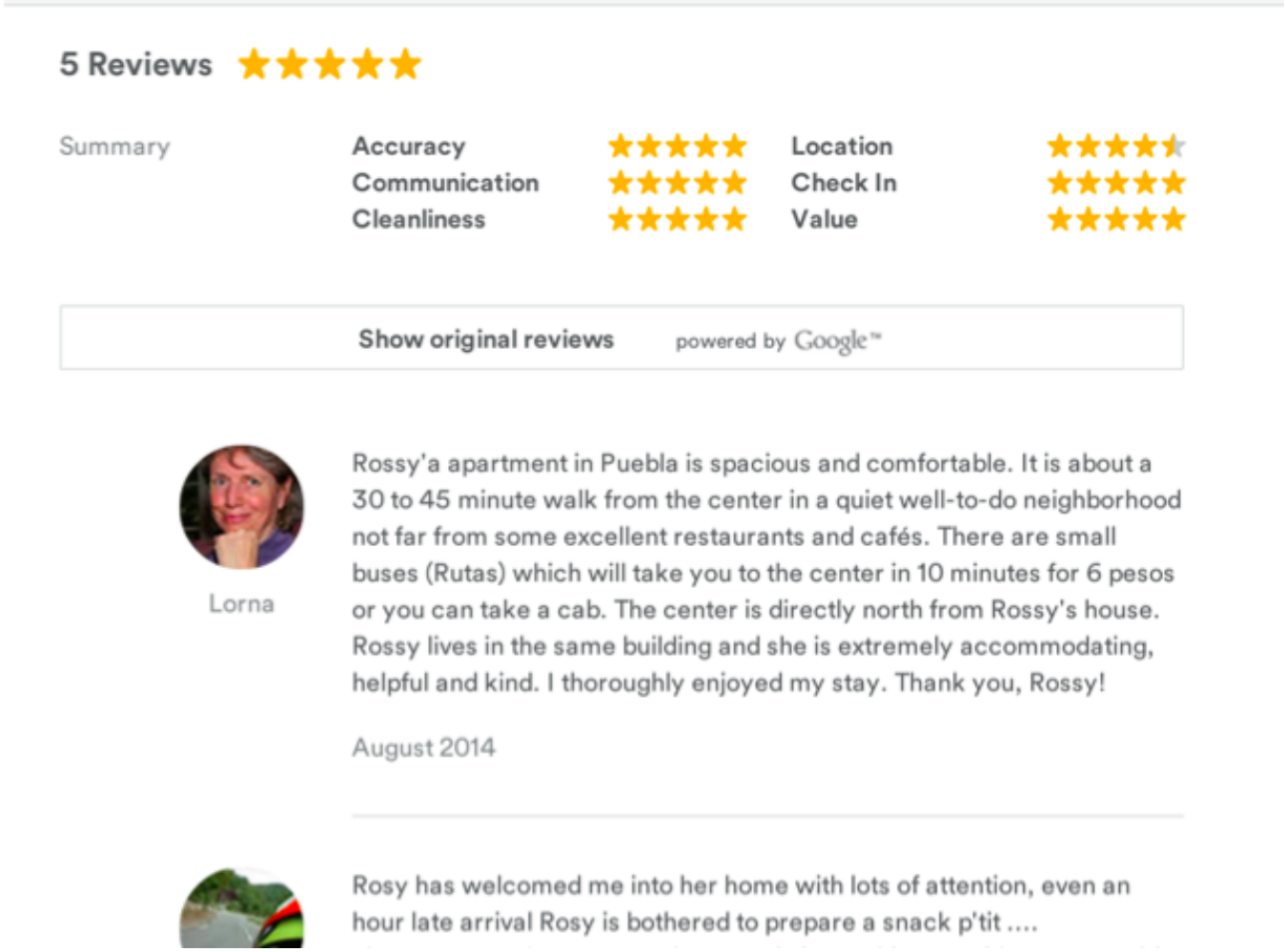


Figure 2: Review of Listing Form (Page 1)

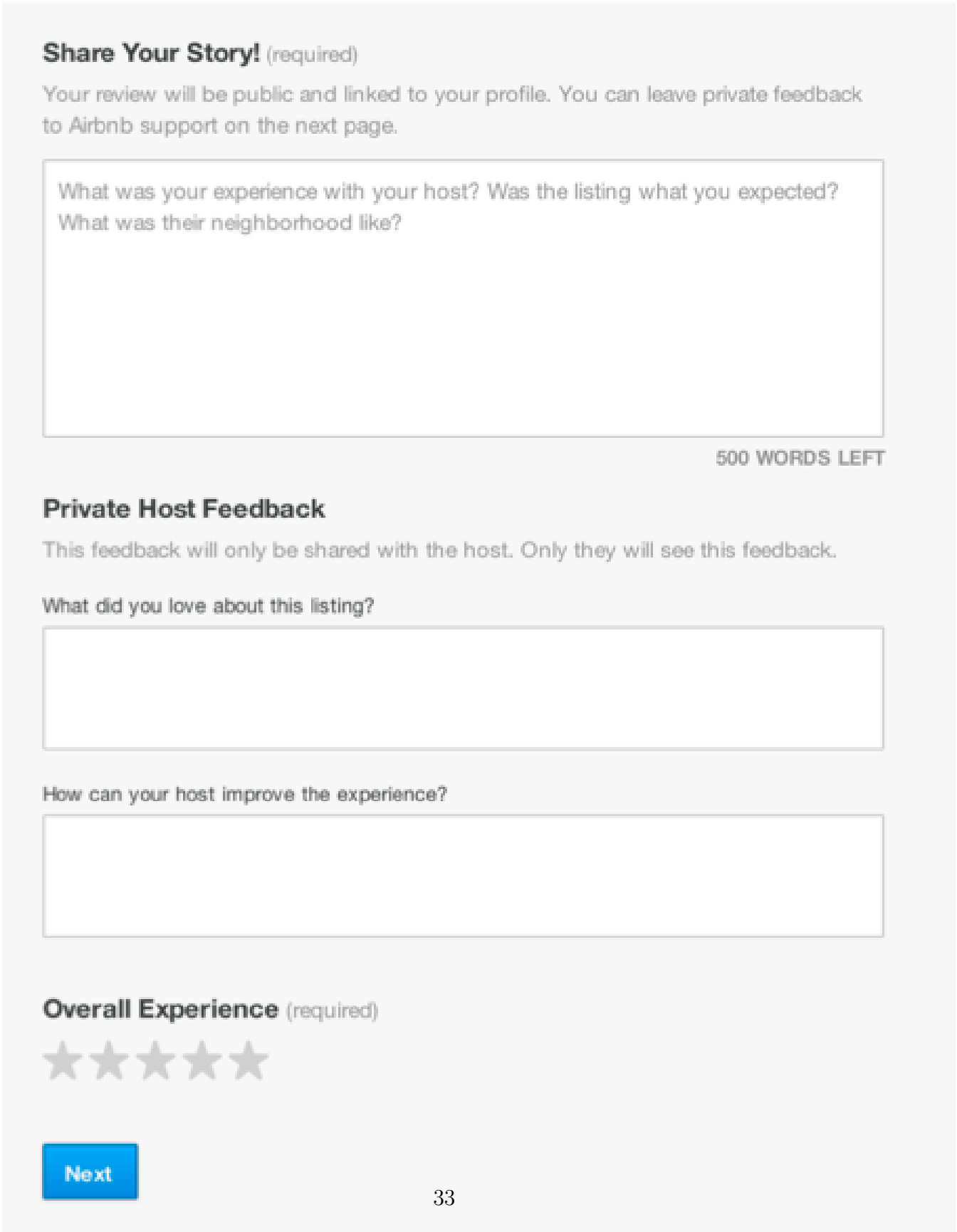


Figure 3: Review of Guest Form (Page 1)

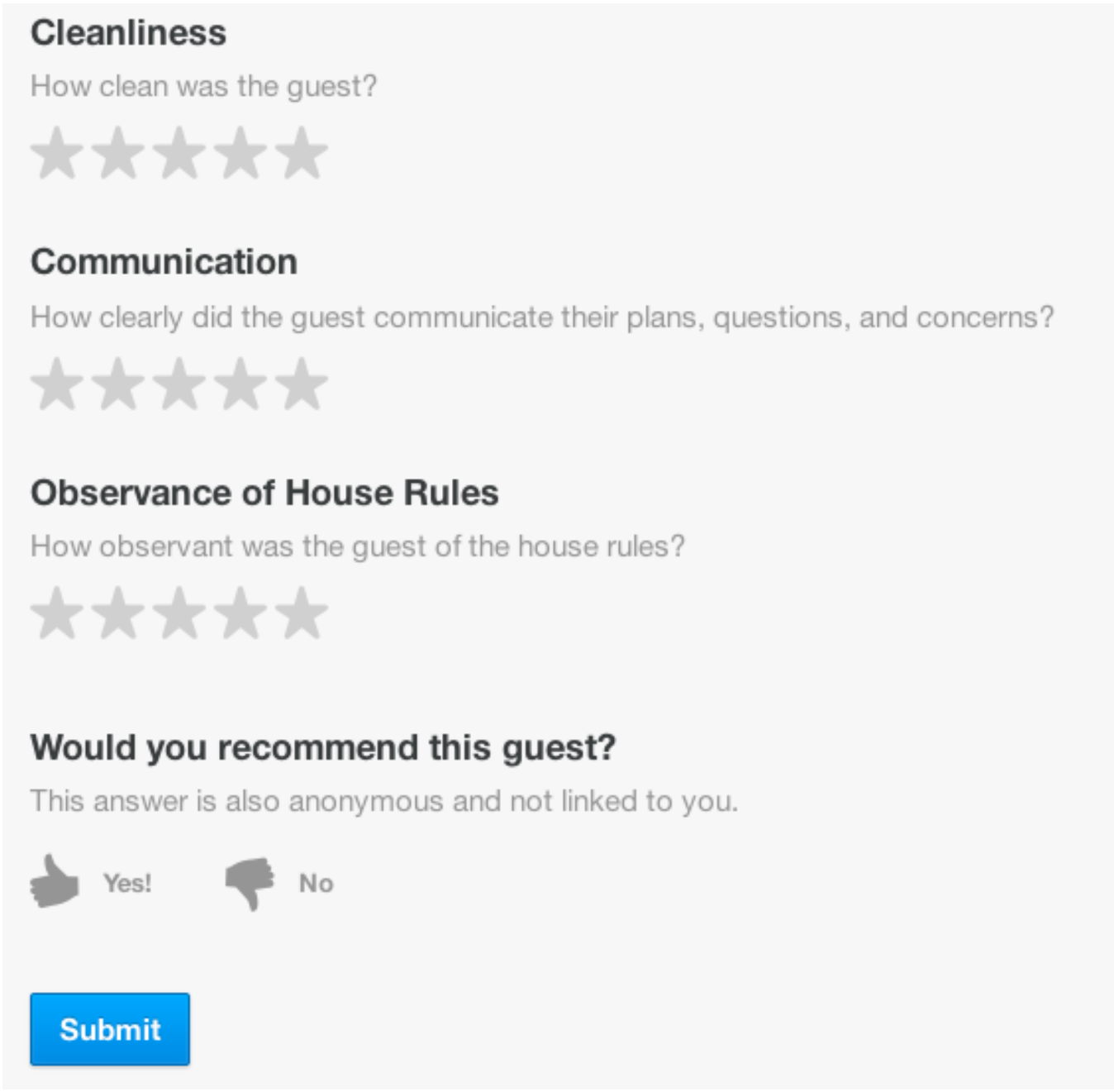


Figure 4: Coupon Experiment - Treatment Email

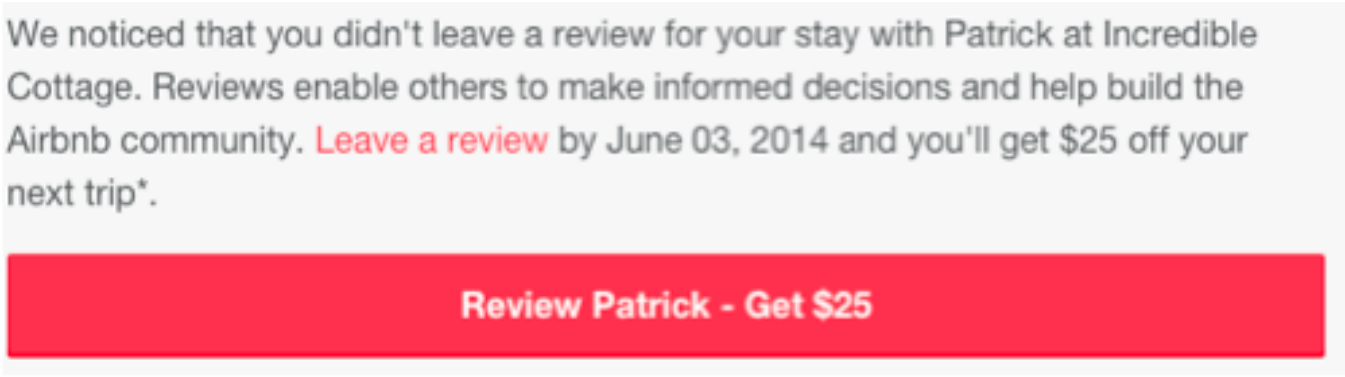


Figure 5: Coupon Experiment - Control Email

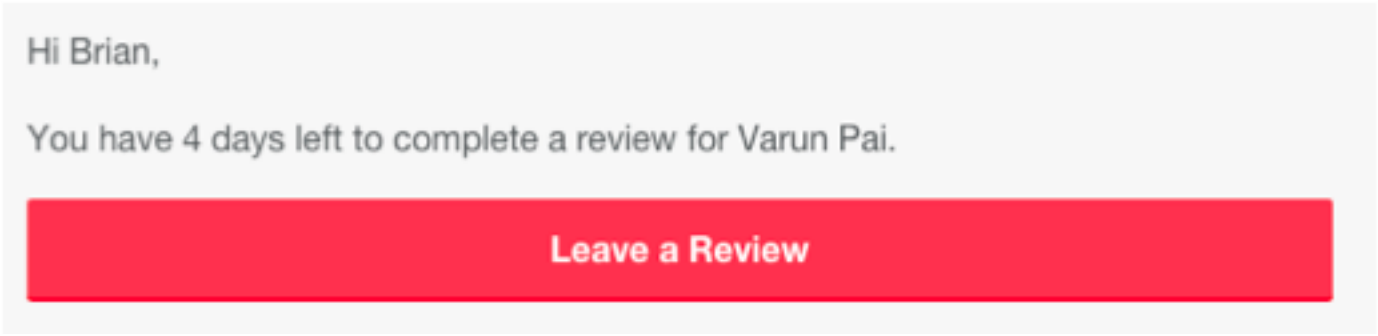
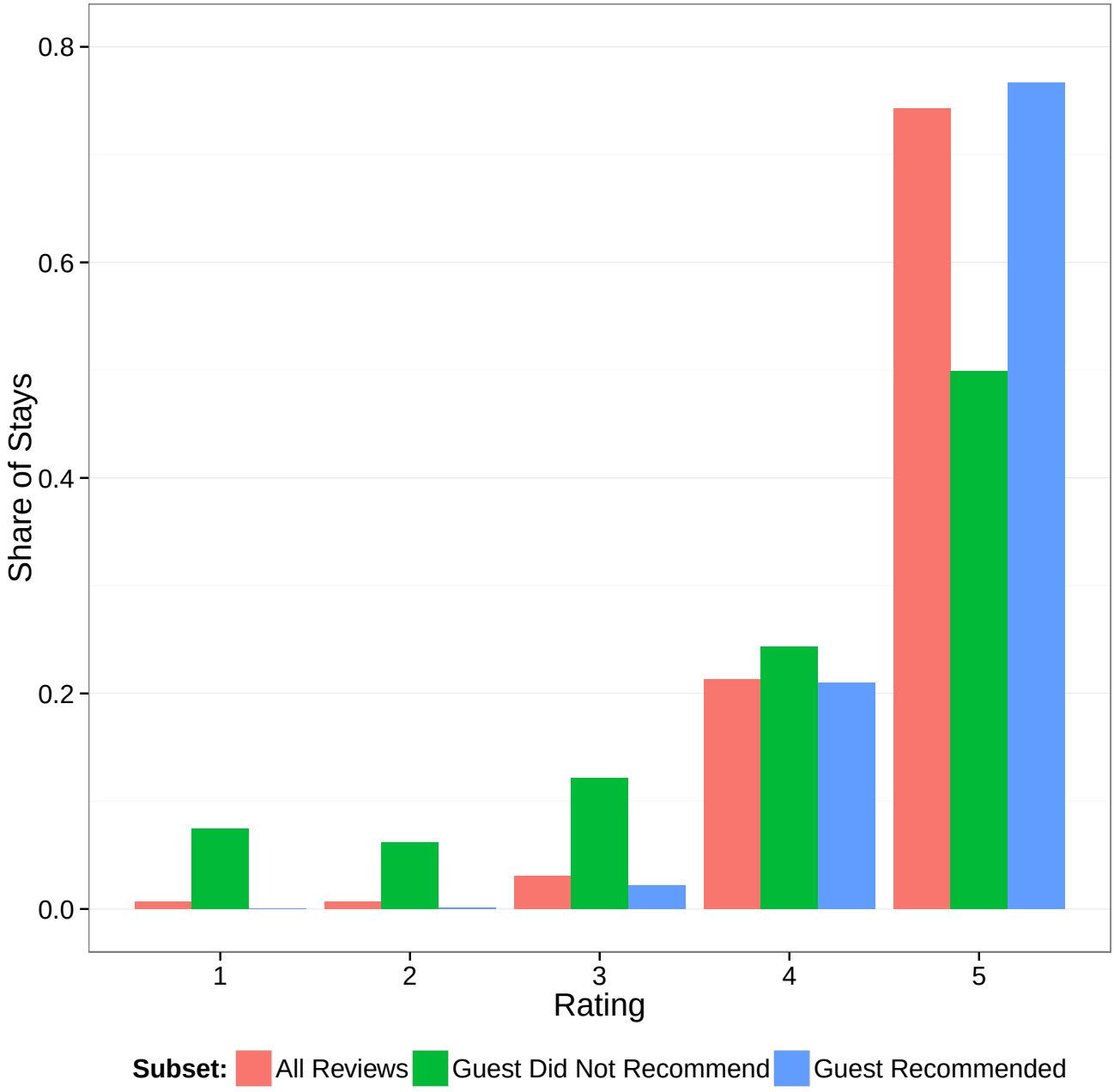
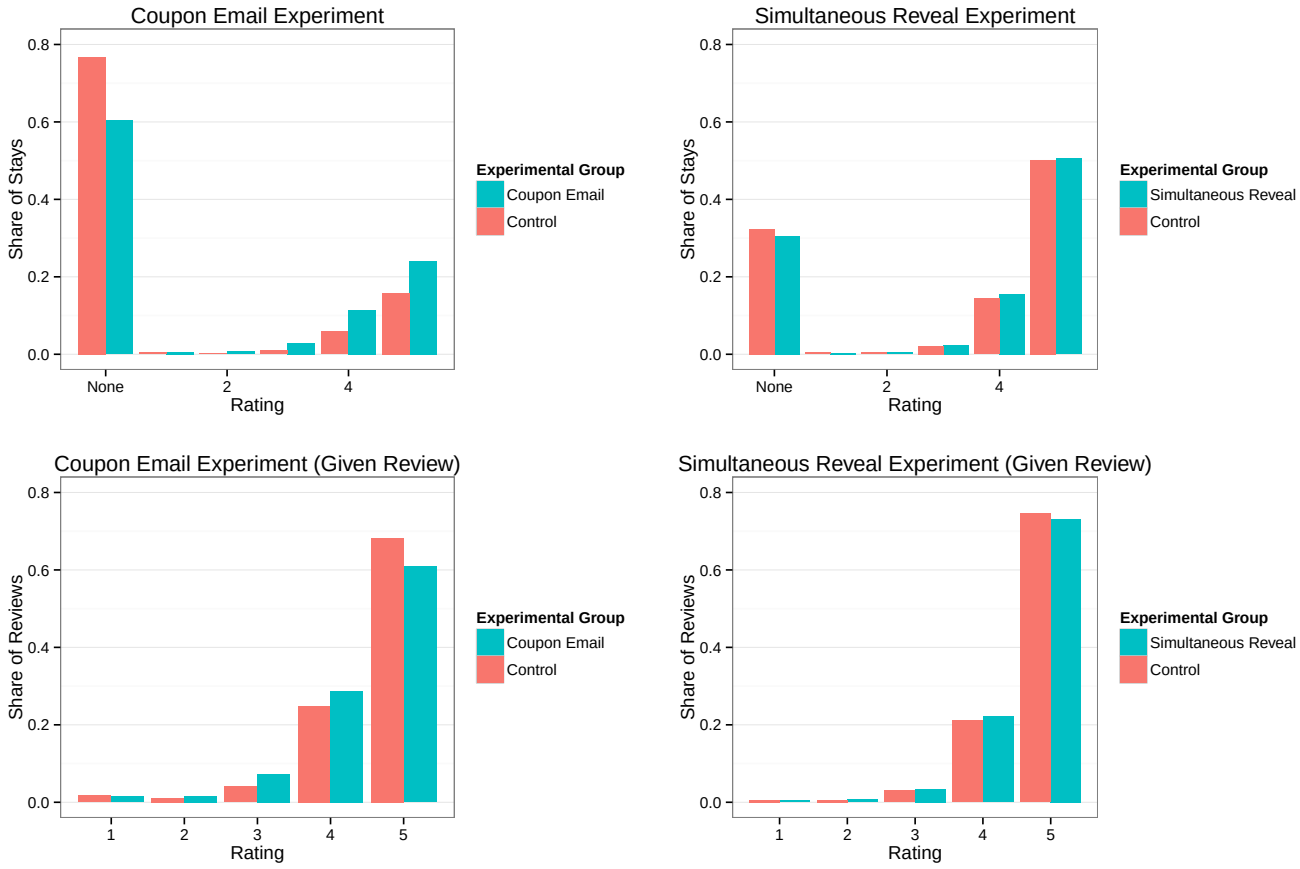


Figure 6: Distribution of Guest Overall Ratings of Listings



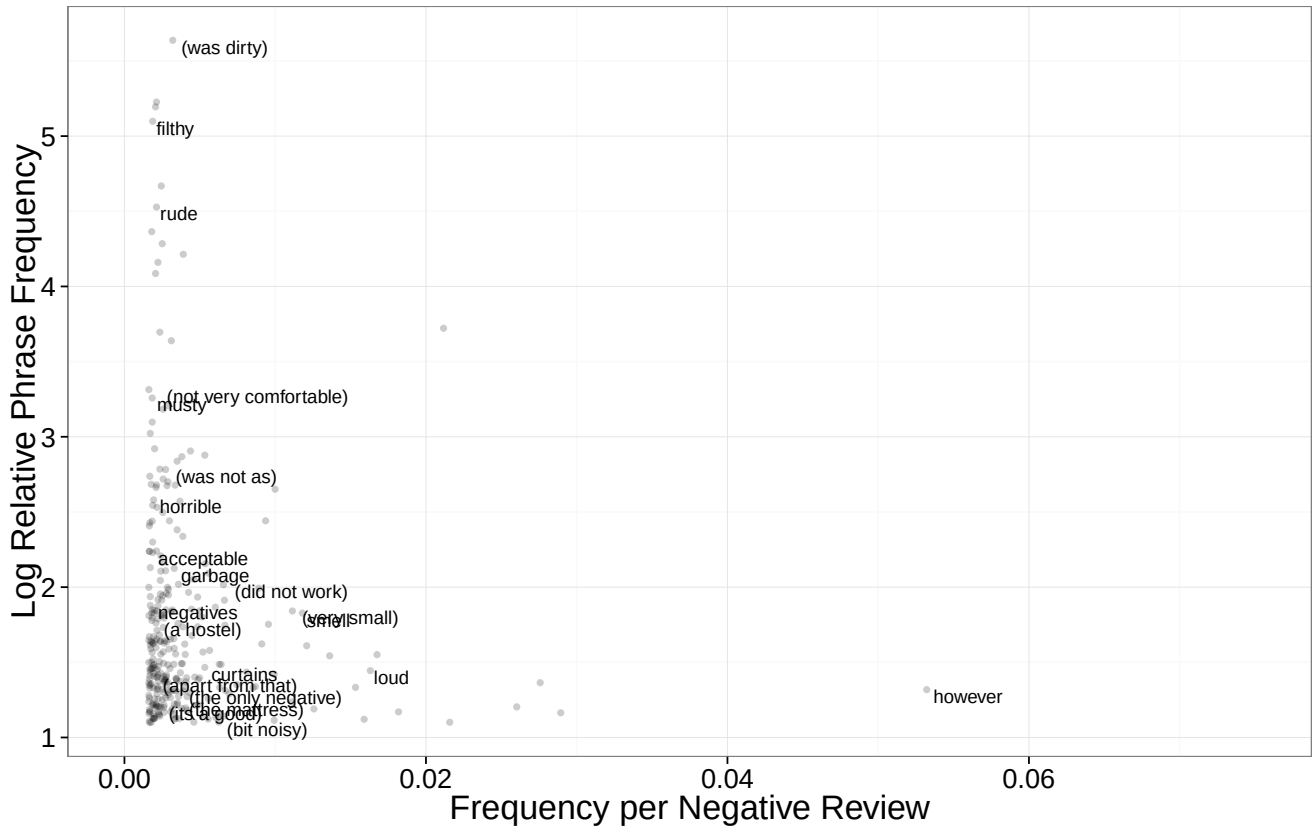
The above figure displays the distribution of submitted ratings in the control group of the simultaneous Reveal experiment (only first stays in the experimental time period are included). "Guest Did Not Recommend" refers to the subsample where the guest responded to an anonymous question that they would not recommend the listing. "Guest Recommended" is the analogous sample for those that did recommend the listing.

Figure 7: Distribution of Ratings - Experiments



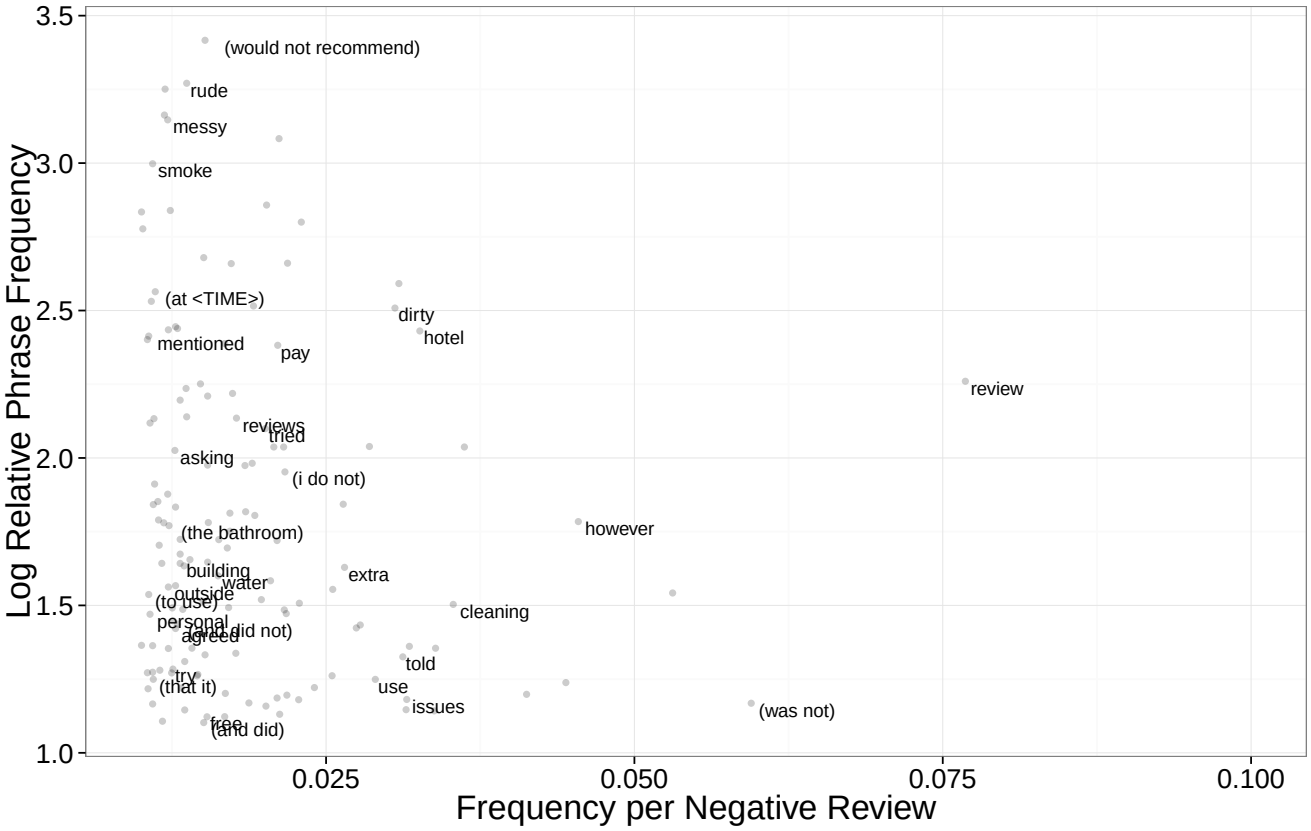
The above figure displays the distribution of ratings in the control and treatment groups for the Simultaneous Reveal Experiment and for the Incentivized Review Experiment. Row 1 displays the distribution of reviews while row 2 displays the distribution of overall ratings.

Figure 8: Phrases Common in Low-rated Reviews by Guests



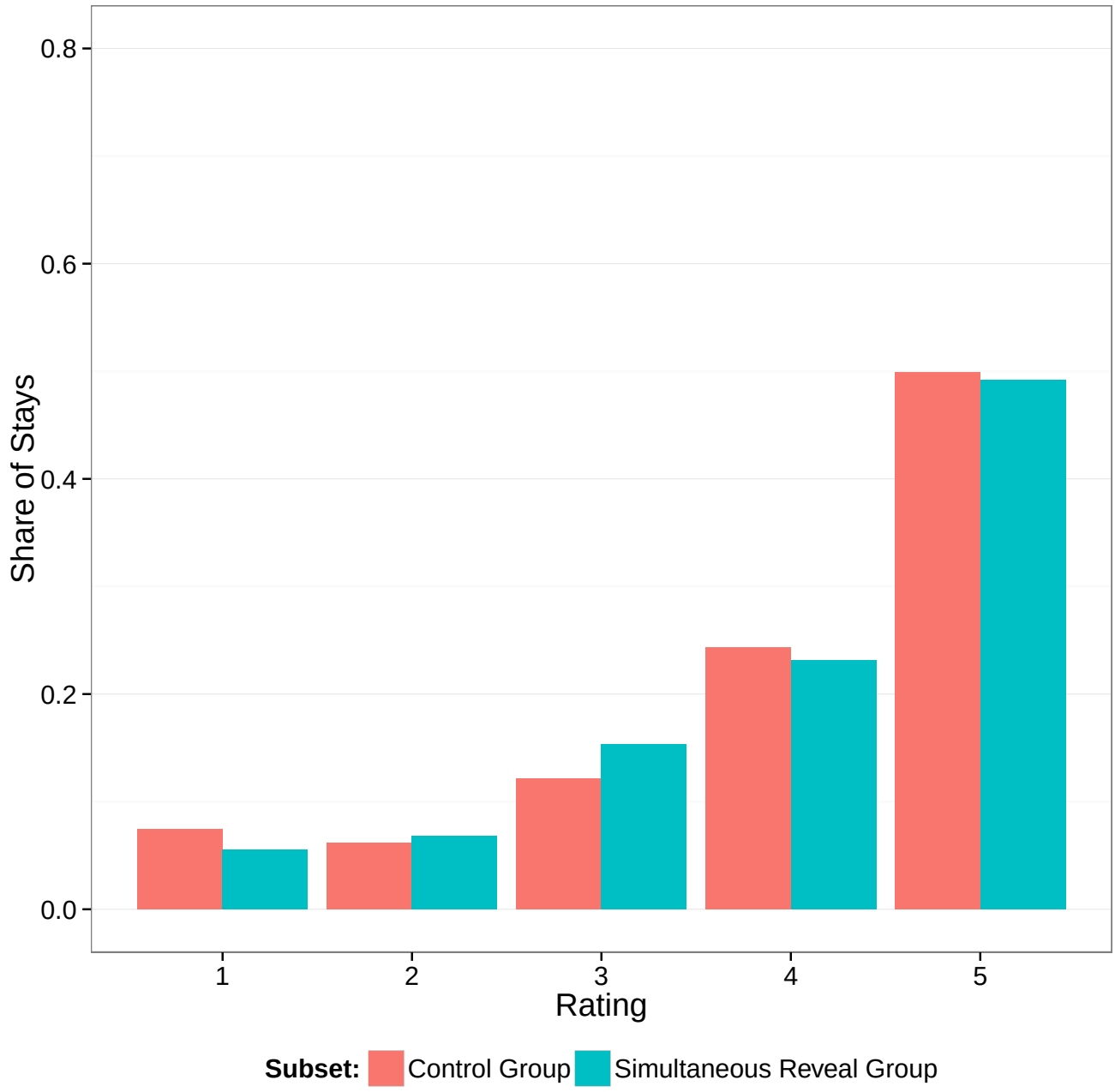
This figure displays all phrases (1, 2, and 3 words) which were at least 4 times as likely to appear in reviews where the guest left a lower than 5 star overall rating than in reviews where the guest left a 5 star rating.

Figure 9: Phrases Common in Low-rated Reviews by Hosts



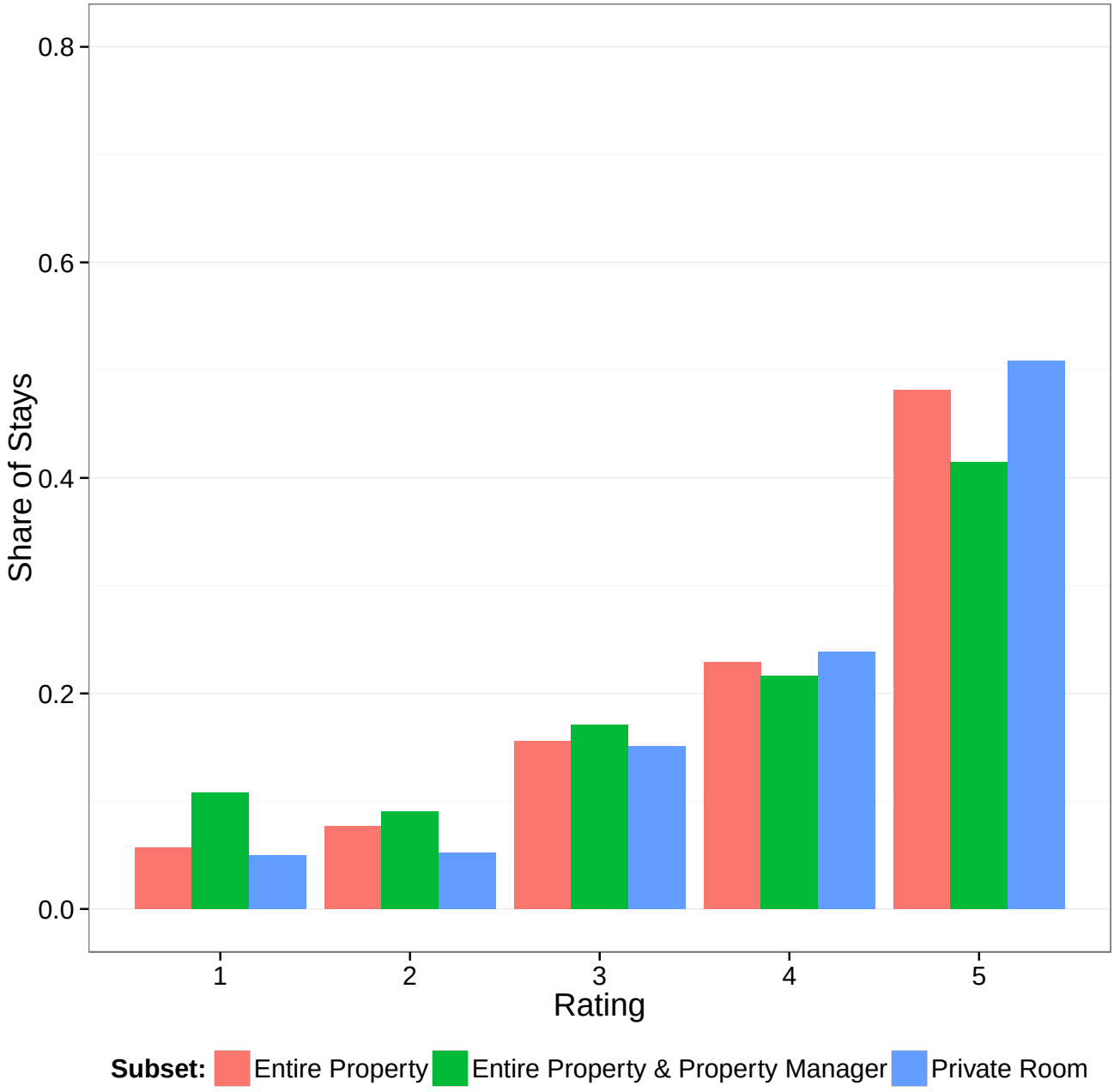
This figure displays all phrases (1, 2, and 3 words) which were at least 4 times as likely to appear in reviews where the host did not recommend the guest than in reviews where the host did recommend the guest.

Figure 10: 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 11: 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	0.913	0.989	0.913	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. *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	0.893	0.986	0.876	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: 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	0.164	0.067	0.024	0.012
Five Star	-0.077	-0.043	-0.016	-0.010
Recommends	-0.021	-0.017	0.000	0.002
Neg. Sentiment	0.011	0.015	0.019	0.027

Columns (1), (2), and (3) display treatment effects in a linear probability model where the dependent variable is listed in the first column. Each regression includes controls for trip and reviewer characteristics: 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. 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. *p<0.10, ** p<0.05, *** p<0.01

Table 6: 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^{**}		
			(0.028)		
Treatment * Guest Judiciousness			-0.025		
			(0.036)		
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. *p<0.10, ** p<0.05, *** p<0.01

Table 7: Retaliation and Induced Reciprocity - Guest

	Does Not Recommend	Overall Rating < 5		Negative Sentiment	
	(1)	(2)	(3)	(4)	(5)
Treatment	0.003	0.030***	0.023	0.032***	0.036*
	(0.004)	(0.006)	(0.025)	(0.005)	(0.021)
Host Negative Sentiment	0.694***	0.615***	0.022	0.377***	0.478
	(0.091)	(0.138)	(0.357)	(0.124)	(0.301)
Host Does Not Recommend	0.079	0.050	0.454	0.254**	0.076
	(0.073)	(0.110)	(0.342)	(0.099)	(0.288)
Treatment * Host Negative Sentiment	−0.706***	−0.609***	0.112	−0.469***	−0.646*
	(0.115)	(0.173)	(0.429)	(0.155)	(0.361)
Treatment * Host Does Not Recommend	0.116	0.245*	−0.323	−0.006	0.245
	(0.093)	(0.140)	(0.394)	(0.126)	(0.332)
Guest, Trip, and Listing Characteristics	Yes	Yes	Yes	Yes	Yes
Only Guest Does Not Recommend	No	No	Yes	No	Yes
Observations	19,361	19,361	1,483	19,361	1,483

The above regressions are estimated for the sample where the host reviews first. "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	0.005	0.003	0.007
	(0.002)	(0.009)	(0.002)	(0.009)
Guest Review Negative	0.053***	0.057***	0.038***	0.034**
	(0.003)	(0.016)	(0.004)	(0.016)
Guest Review Negative Words			0.258***	0.316***
			(0.011)	(0.050)
Guest Does Not Recommend			0.022***	0.002
			(0.006)	(0.026)
Treatment * Review Negative	−0.039***	−0.044**	−0.026***	−0.022
	(0.004)	(0.021)	(0.005)	(0.022)
Treatment * Review Negative Words			−0.207***	−0.280***
			(0.015)	(0.068)
Treatment * Does Not Recommend			−0.016**	−0.012
			(0.008)	(0.037)
Guest, Trip, and Listing Characteristics	Yes	Yes	Yes	Yes
Observations	13,696	7,821	11,107	7,785

The above regressions are estimated for the sample where the guest reviews first. "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***
				(0.010)
Treatment * Does Not Recommend				0.121***
				(0.012)
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. *p<0.10, ** p<0.05, *** p<0.01

Table 10: Fear of Retaliation - Guest

	Reviews First	Not Recommend (First)	< 5 Rating (First)	Neg. Sentiment (First)
	(1)	(2)	(3)	(4)
Treatment	−0.013***	0.002	0.007*	0.010**
	(0.003)	(0.004)	(0.004)	(0.005)
< 5 Rating				0.194***
				(0.007)
Not Recommend				0.133***
				(0.011)
Treatment * < 5 Rating				−0.013
				(0.010)
Treatment * Not Recommend				0.010
				(0.015)
Guest, Trip, and Listing Characteristics	Yes	Yes	Yes	Yes
Observations	118,347	39,177	30,535	30,534

The regressions in columns (2) - (4) are estimated only for cases when the guest reviews first. "Treatment" refers to the simultaneous reveal experiment. *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***	0.015	0.021**
	(0.003)	(0.010)	(0.011)
Customer Support	0.020	0.079***	0.069***
	(0.015)	(0.015)	(0.015)
Guest Recommends		0.101***	0.121***
		(0.008)	(0.008)
Five Star Review			−0.079***
			(0.005)
Recommends * Treatment		0.053***	0.058***
		(0.011)	(0.011)
Five Star * Treatment			−0.015**
			(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. *p<0.10, ** p<0.05, *** p<0.01

Table 12: Socially Induced Reciprocity - Negative Sentiment

	Does Not Recommend	Negative Sentiment		
	(1)	(2)	(3)	(4)
Private Room	0.012***	−0.015***	−0.014**	−0.020***
	(0.004)	(0.006)	(0.006)	(0.006)
Prop. Manager	0.021***	0.029***	0.038***	0.029***
	(0.006)	(0.009)	(0.011)	(0.009)
Guest Does Not Rec.		0.155***	0.150***	0.142***
		(0.010)	(0.009)	(0.008)
Private Room * Does Not Rec.		−0.052***		
		(0.016)		
Prop. Manager * Does Not Rec.		0.069**		
		(0.027)		
Low LTR			0.050***	
			(0.007)	
Private Room * Low LTR			−0.037***	
			(0.012)	
Prop. Manager * Low LTR			−0.002	
			(0.021)	
Comment to Airbnb				0.084***
				(0.008)
Private Room * Comment to Airbnb				−0.006
				(0.014)
Prop. Manager * Comment to Airbnb				0.058**
				(0.026)
Guest, Trip, and Listing Characteristics	Yes	Yes	Yes	Yes
Market FE	Yes	Yes	Yes	Yes
Observations	40,511	31,612	28,173	31,612

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

Table 13: Socially Induced Reciprocity - Star Rating

	Not Five Star		
	(1)	(2)	(3)
Private Room	−0.023***	−0.023***	−0.022***
	(0.006)	(0.006)	(0.006)
Prop. Manager	0.074***	0.075***	0.070***
	(0.009)	(0.011)	(0.009)
Guest Does Not Rec.	0.245***	0.213***	0.246***
	(0.010)	(0.008)	(0.008)
Private Room * Does Not Rec.	−0.001		
	(0.016)		
Prop. Manager * Does Not Rec.	0.036		
	(0.027)		
Low LTR		0.244***	
		(0.007)	
Private Room * Low LTR		−0.022**	
		(0.011)	
Prop. Manager * Low LTR		0.0003	
		(0.020)	
Comment to Airbnb			0.033***
			(0.009)
Private Room * Comment to Airbnb			−0.020
			(0.015)
Prop. Manager * Comment to Airbnb			0.063**
			(0.026)
Guest, Trip, and Listing Characteristics	Yes	Yes	Yes
Market FE	Yes	Yes	Yes
Observations	40,511	36,001	40,511

The outcome in the above regression is whether the guest's overall rating is lower than 5 stars. "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. *p<0.10, ** p<0.05, *** p<0.01

Table 14: Size of Bias (Guest Reviews of Listings)

	Measure of Bias:		
Counterfactual:	B_{avg} Average	B_{mis} % Misreported	B_{neg} % Negative Missing
All Biases	0.078	0.060	0.837
No Strategic Bias	0.076	0.059	0.832
No Social or Strategic Bias	0.063	0.047	0.754
No Social, Strategic or Sorting Bias	0.057	0.057	0.691

The above table displays three measures of bias under four scenarios. 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" further makes the rates of reviews for those with positive and negative experiences to be the same (while keeping the overall review rate constant).