Do Incentives to Review Help the Market? Evidence from a Field Experiment on Airbnb

Andrey Fradkin* David Holtz‡

January 14, 2022

Abstract

Online reviews are typically written by volunteers and, consequently, accurate information about seller quality may be underprovided. We study the extent of this under-provision in a randomized experiment conducted by Airbnb. In the treatment, buyers are offered a coupon to review listings that have no prior reviews. The treatment induces additional reviews, which are more negative on average. Induced reviews do not change nights sold, although they affect the types of transactions that occur. Measures of transaction quality for treated sellers do not improve. We show how market conditions and the design of the reputation system can explain our findings.

*fradkin@bu.edu, Primary Author

†dholtz@haas.berkeley.edu

‡We thank Dean Eckles, Chiara Farronato, Shane Greenstein, John Horton, Caroline Hoxby, Xiang Hui, Ramesh Johari, Garrett Johnson, Jon Levin, Tesary Lin, Mike Luca, Steve Tadelis, Catherine Tucker, Giorgos Zervas, and seminar participants at Microsoft, ACM EC'15, NBER Summer Institute, CODE, WISE, Marketing Science (2021), and the University of Rochester for comments. We thank Elena Grewal and Riley Newman for giving us the initial opportunity to work on this project, Matthew Pearson for early conversations about the project, and Peter Coles and Mike Egesdal for their tireless efforts in helping this paper be approved. The views expressed in this paper are solely the authors' and do not necessarily reflect the views of Airbnb, Inc. The authors were employed by Airbnb, Inc. for part of the time that this paper was written and have held stock that may constitute a material financial position.

1 Introduction

Reputation systems are used by nearly every digital marketplace to help match buyers and sellers. Although reputation systems are considered critical to the success of digital marketplaces, they are known to suffer from a variety of biases and imperfections (Tadelis, 2016). One source of these biases and imperfections is that not everyone reviews. Although in theory increasing review rates should generate additional information and improve markets, we have little empirical evidence as to whether this is the case. We study whether reviews are underprovided through an experimental evaluation of an incentivized review policy on Airbnb.

There are two reasons that incentivized reviews may reduce the harm to markets caused by the underprovision of reviews. First, incentives to review can increase the speed of learning about seller quality by generating more reviews. Acemoglu et al. (2022) show that faster learning about seller quality increases welfare. Second, a non-representative set of reviews may result in the formation of bad matches because buyers may be less aware of negative or polarizing characteristics of a seller. Reviews induced by incentives are more likely to generate information about these characteristics (Dellarocas and Wood (2007), Burtch et al. (2018), Marinescu et al. (2021)). Both of these reasons should be especially relevant for new sellers, since there is high uncertainty about the quality of services they provide (Pallais (2014)).

We provide the first analysis of an incentivized review policy that also measures the effects of incentivized reviews on market outcomes. We find that, although incentives increase review rates, they do not generate economically meaningful benefits to the platform or its users. The experiment we study was conducted by Airbnb between 2014 and 2016 and provided a $25 coupon in exchange for a review of a listing without reviews. In particular, guests to treated listings were sent an email offering them an Airbnb coupon in exchange for a review if they had not reviewed within (typically) 8 or 9 days after checkout, while guests to control listings were instead sent an email reminding them to leave a review.

The incentive increased review rates by 53% and resulted in reviews with lower ratings, on average. Incentivized reviews did not affect the total quantity (nights) sold for treated listings and

caused a change in the composition of transactions. Treated listings had more transactions but they were booked for fewer nights on average. As a consequence the revenue effects of these reviews were not statistically distinguishable from zero.

Incentivized reviews also failed to improve transaction quality and, according to some measures, resulted in worse matches. In particular, the treatment did not affect complaint rates and reduced the post-transaction usage of Airbnb by guests staying with listings after the review. This finding, along with the finding about the lack of revenue effects, suggests that reviews were not underprovided for the treated listings.

Our finding of a lack of benefits from incentivized reviews can be attributed to the market structure and reputation system of Airbnb. If additional reviews arrive quickly after experimental assignment, then the effect of receiving an incentivized review (or not) will be muted. Alternatively, if sellers cannot obtain additional transactions and reviews without an initial review, then the effect of reviews will be large. We show that reviews from other transactions arrive quickly for our experimental sample. While this result is specific to Airbnb, the market structure is not. Using data from another large marketplace, we document a similarly quick rate of review arrivals.

Incentivized reviews may also have small effects on listing outcomes because of Airbnb's reputation system design at the time of the experiment (2014 to 2016). In particular, star ratings (as opposed to the text and number of reviews) were rounded to the nearest half a point and were only displayed once a listing had at least three ratings. As a result, an induced rating was averaged with at least two other ratings when displayed to guests on Airbnb. This averaging and rounding attenuated perceived differences between the ratings of control and treatment listings. We expect similar mechanisms to exist in other platforms which round to half a star including Amazon, Etsy, and Yelp.

An implication of our findings is that institutional details such as market conditions and reputation system design are critical for understanding the role of reviews and the effects of reputation system designs. We do not claim that that reviews and review systems have little value. Indeed, prior work has shown that reputation systems substantially increase consumer surplus (Lewis and

Zervas (2016) and Reimers and Waldfogel (2021). Instead, we show that additional reviews do not help when listings are expected to receive a flow of reviews and when review ratings are displayed as averages.

We contribute to several related research themes within the study of online marketplaces. Because evaluations will be underprovided in equilibrium in the absence of an appropriate payment scheme (Avery, Resnick and Zeckhauser, 1999), a number of recent papers have studied the effectiveness of incentivized review policies and nudges to review at increasing review rates in online settings. Burtch et al. (2018), Marinescu et al. (2021), and Karaman (2020) document that incentives can generate more representative reviews.[1] In a related stream of work, Li (2010), Li and Xiao (2014), Cabral and Li (2015), and Li, Tadelis and Zhou (2020) study policies in which the seller (rather than the platform) offers a rebate for a review.

One existing piece of research that is particularly closely related to our work is Pallais (2014), which experimentally measures the effects of an intervention in which new sellers are both hired and reviewed. She finds that hiring workers and leaving positive feedback has large positive effects on subsequent demand. In contrast, the policy we study generates reviews only for the subset of sellers who are able to transact before receiving their first review.

There is also an emerging research literature that aims to model the dynamics with which consumers learn about seller quality. Our work is closely related to Besbes and Scarsini (2018) and Acemoglu et al. (2022), which both compare consumer learning dynamics under reputation systems in which the full rating history is displayed to learning dynamics under ratings systems in which only summary statistics are displayed. We contribute to these research literatures by experimentally studying the implications of a change to the design of Airbnb's reputation system not only on reviews, but also on subsequent market outcomes.[2]

[1] A preliminary analysis of this experiment was presented in Fradkin et al. (2015). That analysis focused on the first month of data and did not study market outcomes.

[2] Laouénan and Rathelot (2020) and Cui, Li and Zhang (2020) study the effects of Airbnb reviews with a focus on discrimination.

2 Theoretical Framework

Whether the platform should incentivize reviews depends on whether the induced reviews improve outcomes on the platform. We now describe a theoretical framework which clarifies the conditions under which incentivized reviews increase demand and the utility of buyers. Our framework is is a simplified version of Acemoglu et al. (2022), which characterizes the speed of learning in review systems and shows that review systems with a higher speed of learning increase the expected utility of buyers. We derive a formal model of this process in Appendix A.

In our theoretical framework, the degree to which incentivized reviews improve the utility of subsequent buyers is a function of the informativeness of the review system. We conceptualize the informativeness of a review system by the extent to which buyer beliefs about quality after seeing review information (or lack thereof) correspond to the true quality of a listing. This informativeness is a function of both the extent to which ratings correlate with quality and the extent to which buyers' beliefs about reviews correspond to rational expectations. Note that horizontal preferences across listings can be accommodated if buyers first condition on characteristics such as the listing description and photos.

Suppose that a listing has one of two qualities, good or bad, and one of three post-transaction review outcomes: a negative review, no review, and a positive review. Our treatment induces reviews for listings who would have no review were they in the control group.[3] The effects of the treatment on demand are a function of two terms. The first term represents the share of treated listings for which the treatment changes the review outcome from no review to positive review. This is multiplied by the effect on individual demand of one positive review. The demand effect of a positive review corresponds to the change in a buyer's belief that a listing is high quality when a positive review is present.

Similarly, there is a term that represents the share of listings for whom the induced review is negative times the negative effect on demand of a negative review (vs no review). The average

[3]For the purposes of the theoretical framework, we also assume that the coupon offer does not change the reviews of those guests who would have reviewed regardless of the coupon.

treatment effect is the sum of these two terms, which could be positive or negative. Our analysis in section 4 shows that the there is a much bigger increase in positive reviews than in negative reviews. However, it is possible that the increase in demand due to the positive reviews is small and the decrease in demand due to negative reviews is large, in which case the effect of incentivized reviews may be negative.

The strength of any demand effects of one review depend on the extent to which buyer beliefs are updated. Bayesian updating suggests that buyer beliefs about quality should change more when no other reviews are present than when other reviews are present. As a result, the effects of incentivized reviews are mediated by whether the listing is able to quickly obtain other reviews. In subsection 5.3, we document the arrival of reviews from transactions not affected by the experimental incentive.

The effects of incentivized reviews on expected utility in the model are more subtle than the effects on demand. If reviews always corresponded to quality, then incentivized reviews would help buyers identify good and bad listings more quickly. This would increase consumer utility. However, reviews do not perfectly correlate with quality. If incentives cause enough low quality listings to be reviewed positively or enough high quality listings to be reviewed negatively, then the utility of consumers may actually fall due to incentivized reviews. Whether this happens in practice depends on the composition (high or low quality) of non-reviewed listings for whom the incentive induces a review. The net sum of all of these terms is an empirical question which we explore in this paper.

3 Setting and Experimental Design

We analyze an experiment conducted on Airbnb, the largest online marketplace for peer-to-peer short-term accommodations, from April 12, 2014 to May 17, 2016. At the time of the experiment, Airbnb's review system worked as follows. After the guest's checkout, both hosts and guests were asked via email, web notifications, and app notifications to review each other. Both guest and host

reviews consisted of both numeric and textual information. The text of reviews written by guests was displayed on listing pages in reverse chronological order. The numeric overall and category ratings, which were on a one- to five-star scale, were displayed as averages across all transactions rounded to the nearest half star. Rounded average ratings were only visible on listing pages once a lasting had received at least three reviews; before that, only review text was visible on a listing page. The number of reviews was visible on both the search and listing pages as long as the listing had one review. Hence, reviews can have effects both through the search page and through the listing page.

Prior to July 2014, both guest and host reviews were visible both to the counterparty and to the public immediately after submission, and reviews needed to be submitted within 30 days of checkout. Beginning in July 2014, a simultaneous reveal review system was in place (see Fradkin, Grewal and Holtz (2021)). Under the simultaneous reveal system, guests and hosts had 14 days after checkout to submit a review, and reviews were only publicly displayed after both parties submitted a review or 14 days had elapsed. Because our experiment ran from April 2014 to May 2016, the vast majority of our data was collected under the simultaneous reveal review system.

Experiment randomization was conducted at the Airbnb listing level. In order to be eligible for enrollment in the experiment, a listing needed to meet the following criteria:

- It needed to have been booked.
- It needed to have no prior reviews.
- Following a guest checkout, the guest must not have reviewed within a threshold number of days. This number was typically 8 or 9 days throughout most of our sample, with the specific number of days being a function of the email dispatching system. See the Appendix for a more detailed discussion.

Across Airbnb's entire platform, guests who had not reviewed within the threshold number of days described above received an email reminding them to review. For stays at listings that met the criteria above and were assigned to the control group, guests received the standard review reminder

email. For stays at listings that met the criteria above and were assigned to the treatment group, guests received a reminder email that also offered a $25 Airbnb coupon in exchange for leaving a review. These coupons expired one year after being issued, and needed to be used on stays with a minimum cost of $75. Figure B.1 shows the email sent to guests who stayed at treatment listings without reviews during the experiment. In our sample, 326,515 listings were assigned to the control, whereas 328,080 listings were assigned to the treatment. The experiment achieved good balance on pre-treatment covariates and used a well-tested system at Airbnb (Figure C.3).

4 Effects of Experiment on Reviews

In this section, we show that the coupon email treatment induces reviews and that these reviews tend to be worse on average. We first measure the effect of the treatment on the types of ratings and reviews left by guests in either treatment arm who had stays meeting the criteria required to receive the incentive in the treatment group. In particular, we call the first transaction for which a listing is either in the treatment or control the *focal stay*, in contrast to subsequent stays that may also have resulted in reviews. We show that the treatment induced reviews for the focal stay and that those reviews tended to have lower ratings on average. While the textual content of those reviews also tended to contain more negative sentiment on average, there was no difference in text sentiment conditional on the numerical rating.

Figure 1 compares the distributions of numerical ratings left by guests staying at treated listings, and guests staying at control listings. The first thing that is apparent from the figure is that the treatment is effective at increasing the rate at which guests leave reviews: the treatment increases the review rate by 12.9 percentage points, from 24.2% to 37%. Because of this increase in review rate, before conditioning on a review being left, the treatment also increased the number of 5-star reviews (6.3 pp), 4-star reviews (4.6 pp), 3-star reviews (1.4 pp), 2-star reviews (0.31 pp), and 1-star reviews (0.2 pp).

The majority of reviews have a five star rating, which is a fact that has previously been docu-

mented in (Fradkin, Grewal and Holtz, 2021). The high rate of five star reviews has been thought to in part reflect bias in the reputation system of Airbnb. We next measure whether our intervention reduces this bias by changing the distribution of ratings.

The inset in Figure 1 shows that conditional on a review, ratings of treated listings were lower than ratings of control listings; the treatment caused the average rating left by guests to drop by 0.07 stars, from 4.5 to 4.4. The treatment had a lower rate of five star reviews and a higher rate of 2 - 4 star reviews. In other words, while the treatment led to an across-the-board increase in the number of reviews at all levels, the increase was larger for lower ratings than for higher ratings.

We also measure the effects of the treatment on the sentiment of review text. We describe our methodology for text classification and the details of our results about sentiment in subsection B.2. Reviews for treated listings are more negative than for control listings. In particular, 94.1% of reviews in the control group and 93.2% of reviews in the treatment are classified as positive. These differences in text sentiment disappear once we condition on the star rating, which suggests that the effects on star ratings and review text are consistent with each other.

Next, we consider the characteristics of reviewed transactions across treatment and control groups. These characteristics are important since they reveal what types of experiences incentivized reviews reflect. Reviews in the treatment group tend to be lower value, whether measured by nights, bedrooms, or prices (Figure C.4). At the same time, the customer complaint rate does not differ across groups, suggesting that the quality of reviewed transactions is not too different between treatment and control groups.

Considering only treated transactions, reviewed trips are less likely to have customer complaints and have lower transaction values and prices per night (Figure C.5). This suggests that many lower quality transactions (expensive trips and those with customer complaints) are not reviewed even in the treatment.

Our interpretation of the above results is that the incentive induces relatively more of those with lower value and mediocre experiences to review, when they otherwise would not have. The addition of these reviews should reduce the selection bias found in reviews for treated listings. However,

many low quality transactions are not reviewed, even in the treatment. In the next section, we study how induced reviews affect market outcomes.

5 Effects of Incentivized Reviews on Market Outcomes

The platform's objective in incentivizing reviews is to improve market outcomes. In this section, we measure these effects and relate them to our theoretical framework. We begin by showing that reviews have transitory effects on the number of transactions on the platform and that they do not affect overall transaction value or nights booked. We then show that transaction quality is not improved and that, if anything, it falls. Finally, we show how market structure and the design of the reputation system help to explain our findings.

One path that we avoid taking in our main analysis of the treatment's effects is conditioning on whether a review happens or on the rating of the review. Although this approach has intuitive appeal, it suffers from severe omitted variable bias. Whether a listing is reviewed and the rating of the review may be correlated with many other characteristics of the listing which are difficult to control for. These include photos and text, the communication style of the host, the availability of the listing, the quality of competition, and the balance of supply and demand.

5.1 Effects on Demand for a Listing

We begin by measuring the effects of being assigned to the treatment on quantities post-transaction. Our main empirical specifications are linear regressions of the following form:

$$y_{l,h} = \beta_0 + \beta_1 T_l + \epsilon_l \tag{1}$$

where T_l is an indicator for whether the listing, l, had a guest who was sent a treatment email offering a coupon in exchange for the review and $y_{l,h}$ is a listing outcome such as the number of transactions at a time horizon, h. The time horizon encompasses the time between the checkout for

the focal transaction in the experiment and h days afterward.

We consider four listing level outcomes, views after the focal checkout, transactions initiated after the focal checkout, nights for transactions initiated after the focal checkout, and the price to guests of transactions initiated after the focal checkout. Figure 2 displays the results in percent terms. Turning first to views, we see that treated listings receive up to 1% more views, with the effect peaking between around 60 days and then diminishing. Similarly, we see that transactions also increase by about 1% after assignment, with the effect peaking at 120 days. On the other hand, the total nights of stay and booking value exhibit effects close to 0, which are statistically indistinguishable from 0. The effects in percentage terms shrink as the horizon expands, which reflects the temporary effects of the treatment. In subsection B.3 we find that the effect on reservations comes from the intensive margin and that the estimates remain similar when adding controls.

We use a two-stage least squares estimator to translate the effect of a guest receiving an email into the local average treatment effect of an incentivized review. To do so, we must make two assumptions. First, that the coupon email does not change the type of review submitted by those who would have reviewed regardless of the email (the always takers). Second, that the email did not dissuade anyone from reviewing (no defiers). We estimate the following equation:

$$y_{l,h} = \beta_0 + \beta_1 R_l + \epsilon_l \tag{2}$$

where R_l takes the value of 1 if the listing, l, was reviewed for the focal transaction in the experiment and where the instrument is the treatment assignment in the incentivized review experiment. Note that in the case with no covariates, the estimated local average treatment effect of a review will simply scale the estimate in Equation 1 by one over 12.9 percentage points, the causal effect of the coupon email on the review rate.

The second panel of Figure 2 displays the estimated local average treatment effect of an incentivized review. We find that the reviews generate more attention and transactions for listings. Specifically, the effect at 120 days after the focal checkout is 7.1% on views and 9.1% on trans-

actions, which represents an additional 0.33 transactions. Furthermore, the fact that views and transactions increase by similar percentages, suggests that the effect of a review comes from increased clicks from the search page to the listing page. In subsection B.4 we show that this effect is driven by the fact that the number of reviews is displayed on the search page rather than by changes to the algorithmic ranking of a listing.

Even though transactions increase, the number of nights, which represents the total quantity sold, remains constant. This suggests that the presence of an incentivized review changes the *types* of trips that occur. We investigate this by analyzing a transaction level dataset. In particular, we take the set of all post-assignment transactions that are booked within 120 days of assignment and estimate regressions at the trip level, with standard errors clustered at the listing level. We find that nights per trip fall by 1%. See Table C.2 for results on other characteristics, which are statistically insignificant.

To summarize, the net effect of reviews on listing quantity and revenue is close to 0. This is true even though treated listings get more views and transactions. One reason that more transactions do not necessarily translate into more nights is that the capacity of Airbnb listings is limited. In particular, unlike in goods markets, only one buyer can book a listing per night. As a result, increased views can only increase quantity if they result in bookings for marginal nights. Another response by sellers could be to increase nightly prices due to being reviewed but they do not. As shown in Huang (2021), sellers on Airbnb are often inattentive or constrained in changing prices when responding to demand fluctuations.

5.2 Effects on Transaction Quality

Even if incentivized reviews have small effects on average demand, they may affect the quality of matches that occur. This could happen if, for example, the review text contained information that helped potential guests pick listings. To test for this, we construct transaction level customer satisfaction proxies for transactions post-treatment. More concretely, for each listing, l, we consider all transactions that occur within 120 days after the checkout of the focal stay, do not have a

cancelation, and have an observed payment. For this sample of observations, we measure customer complaints, reviews, and customer return rates. Customer return rates are measured by the number of subsequent nights on the platform for guests staying at the listing post-treatment.[4]

Table 1 displays the results of the transaction quality regressions. In particular, the service complaint rate (Column 1) is not statistically different for post-treatment transactions between treated and control listings. The review rate increases, but this is caused in part by the fact that guests of treated listings are eligible for the coupon until the listing has a first review. In particular, for treated listings who do not get a review in the focal transactions, the next guest could also receive an incentivized review offer if they did not review within 7 to 10 days. Column (3) shows that conditional on a subsequent review, the rating is worse in the treated group. This is consistent with the fact that some treated listings may receive an incentivized review from subsequent transactions, rather than the first one and also with a worse match quality in the treatment group.[5] The addition of covariates does not substantively affect these results (Table C.4).

Column (4) of Table 1 displays the effects on the return propensities of guests who stay at listings after the focal transaction. Guests to treated listings stay for fewer nights after a transaction than guests to control listings. The effect is statistically significant and represents a 1.6% decrease in nights.

This effect on guests' subsequent platform usage can be due to one of two mechanisms. The first is that incentivized reviews cause worse matches and cause guests to use the platform less as a result. The second is that induced reviews induce matches with different types of guests, who are also less likely to use the platform afterwards. To investigate this further, in column (5), we add controls for guest and trip characteristics. If different types of guests are induced to stay due to incentivized reviews, then these controls could capture these differences. We still detect a statistically significant effect of the treatment, although the point estimate is smaller.

The small negative effect in column (5) of Table 1 is consistent with a theoretical possibility

[4] User return rates to the platform have been used as a measure of customer satisfaction in Nosko and Tadelis (2015) and Farronato et al. (2020).

[5] Another reason, proposed by Hui et al. (2021), is that low ratings may be autocorrelated due to belief updating dynamics that affect review rates.

that incentivized reviews actually cause *worse* matches. In particular, as described in section 2, incentivized five star reviews may be induced for low-quality listings. As a result, low-quality listings may appear to subsequent guests as higher quality and may thus cause worse matches. We conclude that, if anything, users who stay at treated listings may have a worse experience.

5.3 Why Do Incentivized Reviews Have Small Effects on Demand?

In this section we document two reasons why incentivized reviews have have small effects on demand. The first is that the listings are typically able to generate transactions even without a review. As a result, their other transactions can generate a first review, even if the focal transaction in the experiment did not. These first reviews typically arrive quickly and have a similar ratings distribution to other reviews. The second reason is that star ratings (as opposed to review text) are only displayed as an average after a listing has three reviews. Therefore, differences between treatment and control reviews would be less noticeable by potential guests.

Listings in our sample have been able to get at least one booking without a review. This means that, at least for some guests, the presence of a first review is not pivotal in their choice. One reason that guests take a chance with a non-reviewed listings is that many Airbnb markets are supply-constrained.[6] As a result, guests are shown listings without reviews and sometimes book these listings.

45% of listings in the experiment have more than one booking prior to the checkout of the focal trip. Each of these additional bookings offers an opportunity for the listing to receive a review and these opportunities add up. We find that 72.8% of listings in the control group receive a review by August of 2016, while 78.4% do so in the treatment group. This 5.6% difference is less than half as large as the effect on the treatment for the focal transaction (13%). Furthermore, the control group reviews arrive quickly, the median difference in time between reviews in the control and treatment

[6] We can measure the degree of supply-constraints using the ratio of the number of inquiries divided by the number of listings contacted by market during the time that the experiment was conducted. The average listing in our experiment is booked in a market where the tightness (31.6) is much higher than the tightness in a typical market (18.1).

groups is 6.

Our finding that sellers can quickly obtain other transactions and reviews is not unique to Airbnb. To demonstrate this, we use a scraped dataset from a large home improvement services platform (Farronato et al. (2020)).[7] We find that the median time between a first and second review is 10 days on that platform and that 89% of sellers who have one review manage to get a second. We suspect that other marketplaces have similar dynamics.

Ratings for first reviews occurring outside of the focal stay are more similar between the treatment and control groups than those that occur for the focal transaction. Figure 3 plots the differences in ratings between treatment and control group for reviews coming from the focal transaction and for any first reviews. The effects for first reviews are smaller in magnitude for each rating. This smaller difference in ratings is likely to contribute to the small effects on demand and matching that we find.

Another reason why the effects of incentivized reviews are muted is the manner in which Airbnb displays reviews. Ratings are not shown for every review but are instead averaged and rounded to the nearest half star. Rounding to half a star is a common design online and is used by Amazon, Etsy, and Yelp. Furthermore, while review text is displayed even after the first review, average ratings are only shown after 3 reviews. This reduces the likelihood that one review will change the perception of a guest.

To summarize, even control listings typically obtain first reviews. These other reviews come quickly after experimental assignment and exhibit broadly similar ratings. Furthermore, since ratings are averaged, the effect of just one review is likely to diminish as reviews accumulate. As a result, first incentivized reviews would need to have effects in a relatively short amount of time for them to substantially affect demand and matching outcomes.

[7] Appendix B.5 provides details about this dataset.

5.4 Large Heterogeneous Treatment Effects Do Not Explain the Small Average Effect

Another potential explanation for small average treatment effects is that incentivized reviews have highly heterogeneous effects. Some listings, such as those on the margin of getting additional bookings, may benefit a lot from an incentivized review while others that would have gotten reviewed regardless may primarily face downside risk. We fail to find evidence that large heterogeneous effects drive our main results.[8]

In order to test for heterogeneity with regards to benefits from a review, we need a variable that proxies for the benefit to a listing of a review. One candidate for such a variable is the predicted future demand for a listing. We would expect that a review benefits listings who would have otherwise done poorly on the platform and may not benefit or even hurt listings who are predicted to do well. We construct this proxy in three steps.

First, we select a similar but auxiliary sample on which to train the prediction model. This avoids having to conduct sample splitting procedures as in Guo et al. (2021), who propose a similar way to reduce variance and estimate heterogeneous treatment effects for the purpose of analyzing digital experiments. Our sample consists of previously non-reviewed listings who were reviewed within 9 days of the checkout, and were thus not eligible for our experiment. Intuitively, this is a similar population of listings and so the covariates that predict success on the platform should be similar to those of the experimental sample.

Second, we estimate a linear regression with listing outcomes as a dependent variable and pre-checkout covariates, market, and location fixed effects as control variables. Third, we apply the coefficients from the prior step to the experimental sample in order to create a prediction for each listing in the sample of the listing outcomes.

To test for heterogeneity, we estimate a regression of the following form (as suggested by Lin (2013)):[9]

$$y_l = \beta_0 + \beta_1 T_l + \beta_2 X_l + \beta_3 T_l (X_l - \bar{X}) + \epsilon_l \tag{3}$$

[8] Appendix B.6 shows that large treatment effect heterogeneity by rating does not explain our results either.

In the above regression, y_l is a listing outcome (reservations, nights, and booking value) within 120 days of the focal stay checkout and T_l is the treatment indicator, while X_l is the prediction of the outcomes and $\bar{X}$ is its average. The interaction coefficient, β_3 is our main coefficient of interest.

Table 2 displays the results from Equation 3. Predicted nights are indeed a good proxy since the coefficient on this variable is higher than .5 and the R^2 rises from approximately 0 to between 13% and 19% depending on the regression. Nonetheless, the interaction term is statistically insignificant and small in magnitude. As a result, heterogeneity with regards to potential success on the platform does not explain the small average effects of the treatment.[10]

6 Discussion

We studied when and whether reviews are underprovided by analyzing an experiment in which buyers were incentivized to submit additional reviews. We found that the incentive was successful in inducing reviews. These incentivized reviews exhibited lower ratings, consistent with the presence of selection bias in the reviewing process. Although the treated group was reviewed faster, there were negligible effects on demand and revenue, and potentially negative effects on match quality.

We argue that the structure of the market and the design of the reputation system is critical for understanding our findings. If sellers are are expected to quickly accumulate reviews, then the effect of a marginal review is likely to be small. This effect of incentivized reviews is further reduced by the rounding of ratings to a coarse average. Airbnb exhibits both of these qualities but so do many other marketplaces.

Our negative evaluation of a specific incentivized review program does not preclude other designs from having positive effects. Our treatment induced reviews for a specific set of transactions

[9] Lin (2013) shows that this specification allows $\hat{\beta}_1$ to be consistent for the average treatment effect even in the presence of covariates.

[10] We also conduct a more standard analysis of heterogeneity in Table C.5.

and had imperfect compliance — 37% of treated transactions were reviewed. A policy that induced reviews for a different subset of transactions could have different effects on market outcomes.

Lastly, the incentivized review policy that we study is not well suited toward solving the cold-start problem. In order to solve the cold-start problem, a platform would need to consider alternative interventions. These include subsidizing transactions to sellers without prior transactions, boosting new sellers in search, and hiring 'mystery shoppers' to examine the quality of new inventory. Whether these policies would be successful is an open question that we leave for future work.

References

Acemoglu, Daron, Ali Makhdoumi, Azarakhsh Malekian, and Asuman Ozdaglar. 2022. "Learning From Reviews: The Selection Effect and the Speed of Learning." *Econometrica (Conditionally Accepted)*.

Avery, Christopher, Paul Resnick, and Richard Zeckhauser. 1999. "The market for evaluations." *American economic review*, 89(3): 564–584.

Besbes, Omar, and Marco Scarsini. 2018. "On information distortions in online ratings." *Operations Research*, 66(3): 597–610.

Burtch, Gordon, Yili Hong, Ravi Bapna, and Vladas Griskevicius. 2018. "Stimulating online reviews by combining financial incentives and social norms." *Management Science*, 64(5): 2065–2082.

Cabral, Luis, and Lingfang Li. 2015. "A dollar for your thoughts: Feedback-conditional rebates on eBay." *Management Science*, 61(9): 2052–2063.

Cui, Ruomeng, Jun Li, and Dennis J Zhang. 2020. "Reducing discrimination with reviews in the sharing economy: Evidence from field experiments on Airbnb." *Management Science*, 66(3): 1071–1094.

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.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805*.

Farronato, Chiara, Andrey Fradkin, Bradley Larsen, and Erik Brynjolfsson. 2020. "Consumer protection in an online world: An analysis of occupational licensing." National Bureau of Economic Research.

Fradkin, Andrey, Elena Grewal, and David Holtz. 2021. "Reciprocity and Unveiling in Two-sided Reputation Systems: Evidence from an Experiment on Airbnb." *Marketing Science*.

Fradkin, Andrey, Elena Grewal, Dave Holtz, and Matthew Pearson. 2015. "Bias and Reciprocity in Online Reviews: Evidence From Field Experiments on Airbnb." *Proceedings of the Sixteenth ACM Conference on Economics and Computation*, 641–641.

Guo, Yongyi, Dominic Coey, Mikael Konutgan, Wenting Li, Chris Schoener, and Matt Goldman. 2021. "Machine Learning for Variance Reduction in Online Experiments."

Huang, Yufeng. 2021. "Seller-Pricing Frictions and Platform Remedies."

Hui, Xiang, Tobias J Klein, Konrad Stahl, et al. 2021. "When and Why Do Buyers Rate in Online Markets?" University of Bonn and University of Mannheim, Germany.

Karaman, Hülya. 2020. "Online Review Solicitations Reduce Extremity Bias in Online Review Distributions and Increase Their Representativeness." *Management Science*.

Laouénan, Morgane, and Roland Rathelot. 2020. "Can information reduce ethnic discrimination? Evidence from Airbnb." *American Economic Journal: Applied Economics*.

Lewis, Gregory, and Georgios Zervas. 2016. "The welfare impact of consumer reviews: A case study of the hotel industry." *Unpublished manuscript*.

Li, Lingfang. 2010. "Reputation, trust, and rebates: How online auction markets can improve their feedback mechanisms." *Journal of Economics & Management Strategy*, 19(2): 303–331.

Li, Lingfang, and Erte Xiao. 2014. "Money talks: Rebate mechanisms in reputation system design." *Management Science*, 60(8): 2054–2072.

Li, Lingfang, Steven Tadelis, and Xiaolan Zhou. 2020. "Buying reputation as a signal of quality: Evidence from an online marketplace." *The RAND Journal of Economics*, 51(4): 965–988.

Lin, Winston. 2013. "Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique." *The Annals of Applied Statistics*, 7(1): 295–318.

Marinescu, Ioana, Andrew Chamberlain, Morgan Smart, and Nadav Klein. 2021. "Incentives can reduce bias in online employer reviews." *Journal of Experimental Psychology: Applied*.

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

Nosko, Chris, and Steven Tadelis. 2015. "The limits of reputation in platform markets: An empirical analysis and field experiment." National Bureau of Economic Research.

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

Park, Sungsik, Woochoel Shin, and Jinhong Xie. 2021. "The fateful first consumer review." *Marketing Science*.

Reimers, Imke, and Joel Waldfogel. 2021. "Digitization and pre-purchase information: the causal and welfare impacts of reviews and crowd ratings." *American Economic Review*, 111(6): 1944–71.

Sanh, Victor, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter." *arXiv preprint arXiv:1910.01108*.

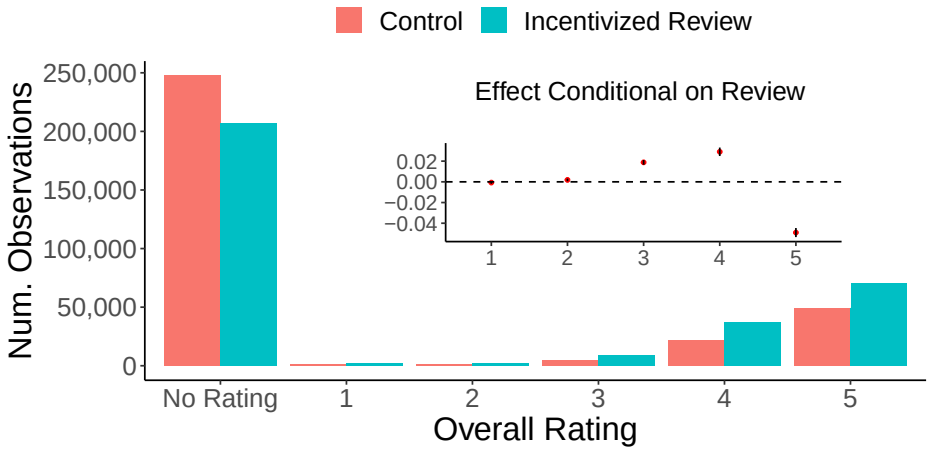
Socher, Richard, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. "Recursive deep models for semantic compositionality over a sentiment treebank." 1631–1642.

Tadelis, Steven. 2016. "Reputation and feedback systems in online platform markets." *Annual Review of Economics*, 8: 321–340.

Wolf, Thomas, Julien Chaumond, Lysandre Debut, Victor Sanh, Clement Delangue, Anthony Moi, Pierric Cistac, Morgan Funtowicz, Joe Davison, Sam Shleifer, et al. 2020. "Transformers: State-of-the-art natural language processing." 38–45.

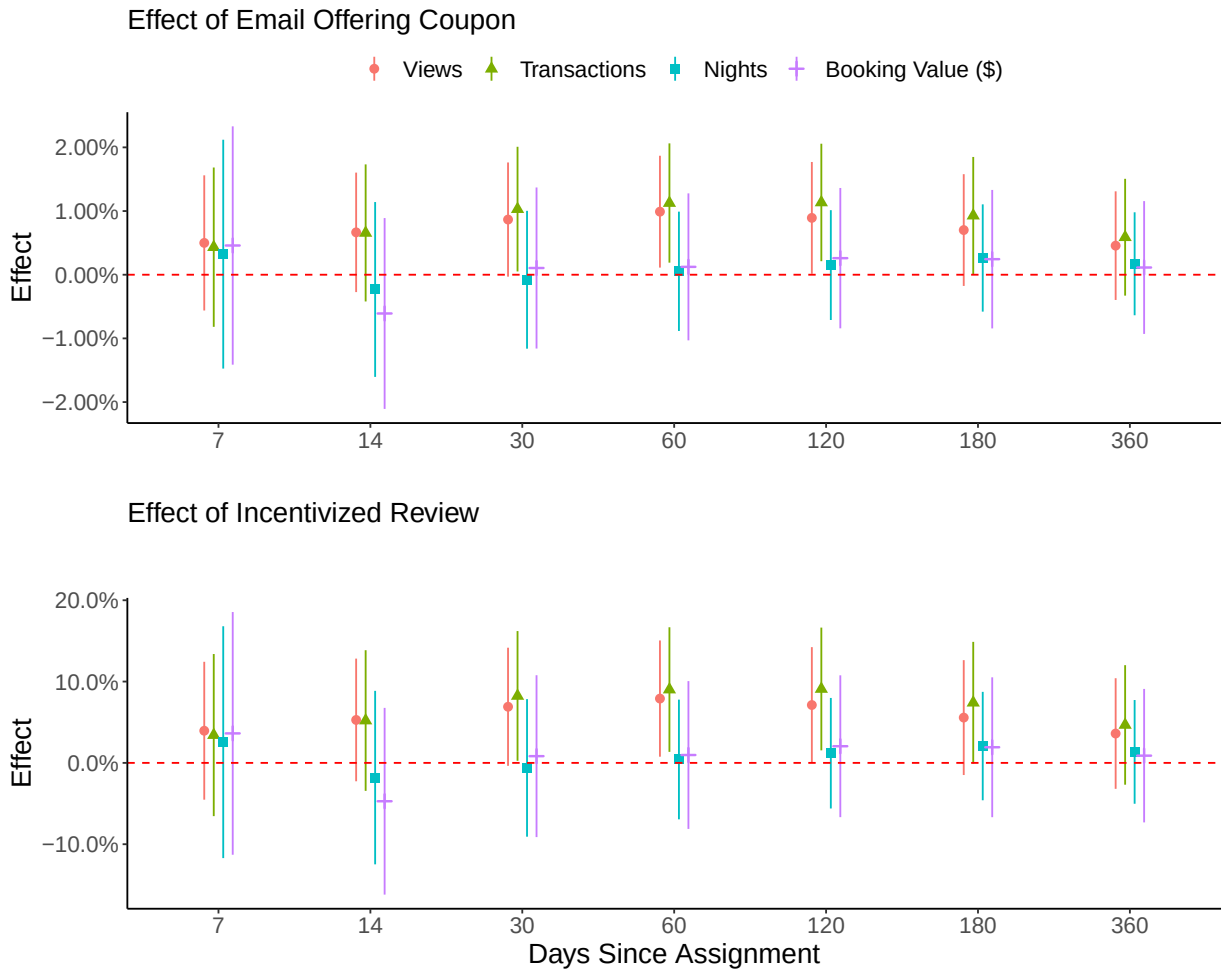
7 Figures and Tables

Figure 1: Distribution of Ratings for Focal Stay



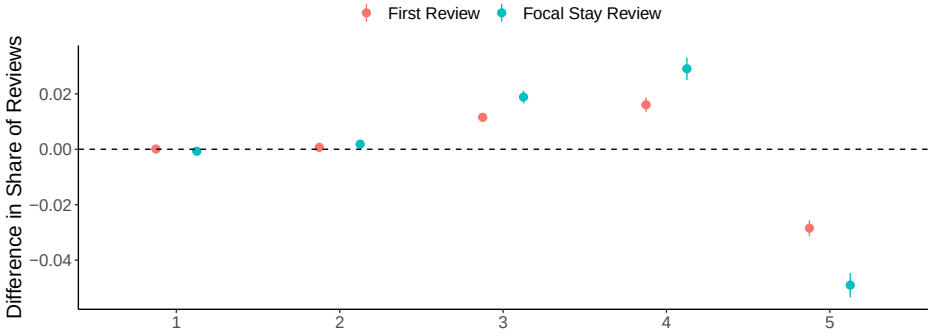
Notes: Comparison of the distribution of ratings left in the treatment group and the control group during the experiment. We only include the first review left for each listing. The inset plot contains the treatment effect and 95% confidence interval conditional on a rating being submitted.

Figure 2: Cumulative Effects of Treatment on Listing Outcomes



Notes: The figure plots the effects and 95% confidence intervals from Equation 1, where coefficients are transformed into percent terms by dividing by the intercept. Each point represents the effect of a listing's guest receiving a treatment email on an outcome measured cumulatively between the checkout for the focal transaction and days since assignment. Standard errors are calculated using robust standard errors and the delta method for the ratio of the treatment coefficient and intercept.

Figure 3: Effect on Review Ratings (Conditional on Review)



Notes: The figure plots the estimate and 95% confidence interval for differences in the share of reviews with each rating type. 'Focal Stay Review' refers to any review that occurred for the first transaction for a listing that was eligible for the experimental treatment. 'First Review' refers to the first review ever received by a listing.

Table 1: Effects of Treatment on Transaction Quality

	Complaint	Reviewed	Star Rating	Guest Nights	
	(1)	(2)	(3)	(4)	(5)
Constant	0.0101***	0.6475***	4.529***	5.591***	
	(0.0001)	(0.0006)	(0.0014)	(0.0217)	
Treatment	-6.52×10^{-5}	0.0048***	-0.0060**	-0.0766**	-0.0548*
	(0.0001)	(0.0009)	(0.0020)	(0.0296)	(0.0245)
R^2	1.06×10^{-7}	2.52×10^{-5}	1.53×10^{-5}	6.48×10^{-6}	0.20805
Observations	2,431,085	2,431,085	1,579,132	2,431,085	2,431,085
Controls	No	No	No	No	Yes
Guest Region FE					✓
Checkout Week FE					✓
Num. Nights FE					✓
Num. Guests FE					✓

Notes: This table displays regressions measuring the effects of the treatment (the guest receiving an email with an offer of a coupon in exchange for a review) on measures of transaction quality. The set of transactions considered for this regression includes all transactions for which the checkout date was between the checkout date of the focal transaction and 360 days after. 'Complaint' refers to whether a guest submitted a customer service complaint to Airbnb, 'Reviewed' refers to whether the guest submitted a review, 'Star Rating' refers to the star rating of any submitted reviews,'Guest Nights' refer to the number of transacted nights for a guest in the 360 days post checkout. Control variables in (5) include the log of transaction amount, the number of times the guest has reviewed and reviewed with a five star ratings in the past, the prior nights of the guest, whether the guest has an about description, and guest age on the platform.

Table 2: Heterogeneity by Predicted Outcomes

	Reservations (1)	Nights (2)	Booking Value (3)
(Intercept)	-0.3024***	0.5944***	91.10***
	(0.0271)	(0.0847)	(17.05)
Treatment	0.0349**	0.0287	4.260
	(0.0152)	(0.0629)	(7.942)
Predicted Reservations	0.5431***		
	(0.0042)		
Treatment × Predicted Reservations (Demeaned)	0.0053		
	(0.0066)		
Predicted Nights		0.5970***	
		(0.0039)	
Treatment × Predicted Nights (Demeaned)		0.0093	
		(0.0062)	
Predicted Booking Value			0.6630***
			(0.0082)
Treatment × Predicted Booking Value (Demeaned)			0.0021
			(0.0102)
Horizon	120 Days	120 Days	120 Days
Observations	640,936	640,936	640,936
R^2	0.16055	0.13454	0.18840

Notes: This table displays the regression estimates from Equation 3, where the outcome is reservations, nights, and booking value within 120 days of the focal checkout. Predicted reservations, nights, and booking values are calculated using the procedure described in subsection 5.4. Note that the number of observations in this regression is lower than in the others since some uncommon fixed effect values in the experimental data were not present in the training data and some covariates were missing for some of the observations.

A Appendix: Theoretical Model

Whether the platform should incentivize reviews depends on whether these reviews improve outcomes on the platform. In this section, we describe a theoretical framework which clarifies the conditions under which incentivized reviews increase demand and the utility of buyers. The framework is a simplified version of Acemoglu et al. (2022), which characterizes the speed of learning in review systems and shows that review systems with a higher speed of learning increase the expected utility of buyers. In this theoretical framework, the degree to which incentivized reviews improve buyer utilities is a function of the informativeness of the review system, which is a measure of the extent to which buyer beliefs about quality correspond to the true quality of a listing. This informativeness is a function of both the extent to which ratings correlate with quality and the extent to which buyer's beliefs about ratings correspond to rational expectations.

Suppose that a buyer is randomly matched with a seller. The seller has a true underlying quality $Q \in \{0, 1\}$ and an associated review outcome $r \in \{-1, 0, 1\}$, where -1 corresponds to a negative review, 0 to no review, and 1 to a positive review. The utility of buyer, i, for listing, l, is:

$$u_{il} = \theta_i + Q_l - p \tag{4}$$

In the above equation $\theta \sim F$ is the ex-ante preference of the buyer for the inside option and p is the price of the listing, which we assume to be constant. The buyer does not know the true value of Q_l and must therefore form a guess based on the review (or lack thereof) and prior beliefs.

The platform has a review system, $\mathbf{\Omega}$, which maps the history of transactions to reviews. Examples of $\mathbf{\Omega}$ include a review system without incentivized reviews and a review system with incentivized reviews. Let Ω_l be a realized outcome of the review system for listing l, where prior buyers have the opportunity to submit reviews. The buyer observes Ω_l and forms a belief $q_i(\Omega_l)$ about the probability that listing q has quality equal to 1. The buyer then makes a utility maximizing purchase decision:

$$b_{il} = \arg\max_{b \in \{0,1\}} \mathbf{1}\{b = 1\}(E_Q[\theta_i + Q_l - p|\Omega_l]) = \arg\max_{b \in \{0,1\}} \mathbf{1}\{b = 1\}(\theta_i + q_i(\Omega_l) - p) \tag{5}$$

Acemoglu et al. (2022) show that in setups similar to this, if consumers have rational expectations and play a pure-strategy Bayesian equilibrium, the beliefs of a sequence of arriving buyers converge to the true seller quality.[11] One boundary condition of this model is worth highlighting. If the upper bound on θ is insufficiently high, then high quality listings may get unlucky. If, for example, $\bar{\theta} < p - q(-1)$, then negatively reviewed listings will never be booked again. This will be the case even if some of those listings are of high quality and were negatively reviewed just by chance. Consequently, this model can allow for results similar to Park, Shin and Xie (2021), where

first negative reviews have large negative effects.

Buyers' expected utilities (across preferences, quality, and realizations of the review system) can be expressed as follows, where we also assume that $\theta \in [p, 1]$ so that people prefer to purchase high quality listings but not low quality listings, μ is the share of listings that are of high quality, and that the belief function, q_i is constant across buyers:[12]

$$\begin{aligned} E_{\theta,Q,\Omega} =& \mu(1 - p + E_\theta[\theta]) \\ &+(1-\mu)E_\theta[-(p-\theta)P_\Omega[q \geq p - \theta | Q = 0]] \\ &+\mu E_\theta[-(1-p+\theta)P_\Omega[q \leq p - \theta | Q = 1]] \end{aligned} \tag{6}$$

The above equation contains the key ingredients necessary for understanding the effects of a change in the reputation system. Line 1 is the utility if everyone only purchased from high quality listings. Line 2 is the false positive utility, which represents the utility loss from purchasing from a low quality listing. Line 3 is the utility lost due to false negatives, which occur when buyers do not purchase from a high quality listing.

We now consider the effects of incentivized reviews on demand and utility in this framework. Suppose that $\mathbf{\Omega}_c$ is the control review system and $\mathbf{\Omega}_t$ is the treatment review system, and further suppose that for any stay, $\mathbf{\Omega}_t$ weakly increases review rates, but results in the same rating conditional on a review. This rules out situations where, for example, the coupon offer changes the degree of reciprocity felt by the guest. We also assume that $q(-1) < q(0) < q(1)$, meaning that positive reviews are better than no reviews and that no reviews are better than negative reviews.

Then the change in demand due to a shift from $\mathbf{\Omega}_c$ to $\mathbf{\Omega}_t$ is:

$$\begin{aligned} &(\tau_{H,1} + \tau_{L,1})Pr(p - q(0) > \theta > p - q(1)) - \\ &(\tau_{H,-1} + \tau_{L,-1})Pr(p - q(-1) > \theta > p - q(0)) \end{aligned} \tag{7}$$

Equation 7 contains two lines. The first line is the increase in demand due to some listings having a positive review in the treatment, where $\mathbf{\Omega}_c(s) = 0$ and $\mathbf{\Omega}_t(s) = 1$. The mass of these listings is $\tau_{H,1} + \tau_{L,1}$, where H and L represent high and low quality listings respectively. This sum is identified in our experiment. For example, the number of five-star reviews increases by 6.3 pp. This sum is multiplied by the change in demand due to a positive review, which is the share of guests that would purchase if the review was high but would not purchase if there were no review. The second line of Equation 7 is analogous but measures the decrease in demand for listings that would have had no review in the control review system but were negatively reviewed

[11] Acemoglu et al. (2022) also place restrictions on the reviewing behavior of buyers.

[12] See the proof of Proposition 6 in Acemoglu et al. (2022) for a more general formulation of this result.

in the treatment system. Our analysis in section 4 shows that the there is a much bigger increase in positive reviews than in negative reviews. However, it is possible that the change in demand due to the positive reviews is small, in which case the effect of incentivized reviews may be small (or negative).

The effects of incentivized reviews on expected utility are a bit more subtle. It could be the case that incentives induce the wrong types of reviews, leading to worse matches. The change in expected utility from incentivized reviews can be expressed as follows:

$$\begin{aligned}
&\tau_{H,1}E[(1-p+\theta)Pr(p-q(0)>\theta>p-q(1))+\\
&\tau_{L,1}E[(-p+\theta)Pr(p-q(0)>\theta>p-q(1))+\\
&\tau_{H,-1}E[-(1-p+\theta)Pr(p-q(-1)>\theta>p-q(0))+\\
&\tau_{L,-1}E[-(p-\theta)Pr(p-q(-1)>\theta>p-q(0))]
\end{aligned} \tag{8}$$

Equation 8 contains four terms, corresponding to cases when high and low quality listings are reviewed either positively or negatively due to the treatment. The best case scenario for an incentivized review system is when the second and third lines are equal to 0, meaning that incentivized reviews increase positive review only for high quality listings and increase negative reviews only for low quality listings. But it may also be the case that incentivized reviews induce positive reviews for low quality listings. This may occur if, for example, guests value the coupon but do not want to say something negative about their stay in a review. In that case, the second line the equation would become relevant.[13] Finally, it may be the case that a high quality listing is unlucky and gets negatively reviewed due to the treatment, a mechanism hinted at in Park, Shin and Xie (2021). That would correspond to line 3.

[13] A similar mechanism is documented in Muchnik, Aral and Taylor (2013), who show that randomly assigned up-votes on Reddit had large positive effects on subsequent scores.

B Appendix: Additional Results

B.1 Description of Review Email Dispatch During the Experiment

The number of days between the checkout and emails in the experiment was intended to be nine days for most of the sample. After March 29 of 2016, the number of days within which a review must have been submitted to determine eligibility was changed to seven.

In practice, the number of days varied for several reasons. First, since transactions happen around the world, the measurement of the date of the checkout and email depends on the time zone in which a checkout occurs. The email system does not perfectly take these time-zones into account. Second, at least during the period we study, stays that had partial cancelations were not fully accounted for by the email dispatch system. As an example, let's say a stay was initially booked for ten days but the guest checked out five days early. The email dispatch system still used the initial ten day booking as the basis for calculating the date of the required email. Third, the exact time of the email varied over time and across days of the week. Lastly, there seemed to be several outages of the email system during which emails were sent with a delay.

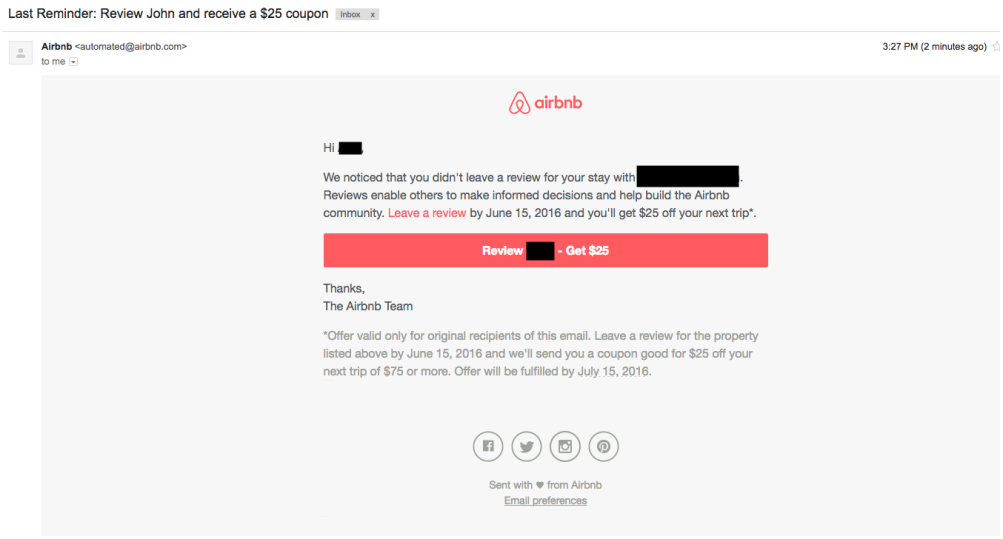
Figure C.6 displays histograms of days between when a listing was assigned to be reviewed by the review system and the date of the email. We can see that prior to April of 2016, the vast majority of emails were sent either 8 or 9 days after checkout. After March of 2016, most emails were sent 7 days after checkout. Figure C.7 plots similar figures where instead the time between the true checkout (accounting for cancellations) and the email is plotted. We see that the days are more dispersed but that the pattern of time to email is similar.

We also measure differences in the days between the checkout and email across the treatment and control groups. On average, emails sent in the treatment arrived 34 minutes later after checkout than emails in the control group. This difference is statistically significant although not economically meaningful. We do not know the exact reason for this difference but suspect it has something to do with the way in which the email dispatch system batched emails. In practice, the treatment can only affect outcomes through inducing additional reviews, and this will occurs even if emails arrive at slightly different times between the treatment and control group.

B.2 Effects on Textual Reviews

In order to measure changes in the textual content of the reviews left by guests, we estimated the sentiment of each review in our sample using DistilBERT (Sanh et al., 2019), which is a lightweight version of BERT, a widely used language model (Devlin et al., 2018). At a high-level, BERT is a model that first pre-trains embedding-based language representations using both the left and right context around words. These pre-trained representations can then be fine-tuned to create models

Figure B.1: Treatment Email



Notes: Displays the email sent to guests who had stayed in treatment listings who had not yet received a review on Airbnb after a certain number of days, inviting them to leave a review in exchange for a coupon.

for a wide variety of natural language processing tasks, such as question answering, language inference, and sentiment analysis. We estimate the sentiment of each review in our sample using the default distilBERT sentiment transformer provided by Huggingface (Wolf et al., 2020), which has been fine-tuned on Version 2 of the Stanford Sentiment Treebank (Socher et al., 2013), a sentiment analysis training set consisting of 11,855 sentences taken from movie reviews.

We find that treated reviews are less likely to have text classified as positive. In particular, 94.1% of reviews in the control group and 93.2% of reviews in the treatment are classified as positive ($p < 3.9 \times 10^{-9}$). Treated reviews are also 8% shorter in length than control reviews.

To investigate whether the changes in review text are consistent with the changes in the star ratings, we regress the text sentiment on indicators for the treatment and the star rating. In particular, we run a regression of the following form:

$$text\_pos_l = \beta_0 + \beta_1 T_l + \gamma_r + \epsilon_l \tag{9}$$

where $text\_pos_l$ is an indicator for whether review text is classified as positive, T_l is a treatment indicator, and γ_r are star rating fixed effects.

Table B.1 displays the results of Equation 9. Column 1 shows that review text in the treatment is less likely to be classified as positive. Column 2 shows that conditional on star ratings, review text is similar between treatment and control listings. Column 2 also shows that star ratings are highly correlated with text sentiment. Reviews with a one star rating have positive text less than 10% of the time while reviews with a five star rating have positive text more than 99% of the time.

Table B.1: Text Sentiment Conditional on Rating

	Text Sentiment Positive	
	(1)	(2)
Constant	0.9405***	0.0951***
	(0.0010)	(0.0062)
Treatment	-0.0080***	-0.0016
	(0.0013)	(0.0010)
2 Stars		0.1677***
		(0.0106)
3 Stars		0.6122***
		(0.0079)
4 Stars		0.8602***
		(0.0062)
5 Stars		0.8979***
		(0.0062)
R^2	0.00026	0.42832
Observations	135,670	135,670

Notes: This table plots regressions results where the outcome is the classified sentiment of the review text and the controls include a treatment indicator and star rating fixed effects.

As a result, we conclude that incentivized reviews differ from regular reviews in similar ways whether measured by text or by rating.

B.3 Additional Analysis of the Effects of Treatment on Listing Outcomes

In this section, we conduct additional analysis of our experimental results. In particular, we investigate whether adding controls substantially effects the precision of our estimates, whether the effects of the treatment on reservations come from the intensive or the extensive margin, and whether hosts adjust their behavior in response to the treatment.

In Table C.1 we display the results of the intent to treat regressions with a 120 day time horizon, with control variables for listing, guest, and focal transaction characteristics. In particular, we control for room type, capacity, bedrooms, prior nights, prior bookings, trips in process, number of listings managed by the host, main photo size, number of photos, guest gender, whether the guest is a host, guest prior nights, guest prior reviews submitted, guest prior five star reviews submitted are included along with checkout week and zip code fixed effects. With these covariates, we detect effects on views and reservations, but not on nights and booking value. This mirrors the results without control variables.

Next, we consider whether the effects on reservations come from the intensive or the extensive margin. Induced reviews may help some listings who would've otherwise failed on the platform or the may hurt some listings with a negative review. For both of these cases, we would expect to see

an effect on the extensive margin, i.e. whether a listing gets subsequent reservations. On the other hand, if induced reviews affect the types and frequency of subsequent transactions, then this effect may be felt on the intensive margin.

In Table C.3, we estimate separate regressions where the outcome is whether a listing has a reservation at all, and how many reservations a listing has conditional on receiving at least one subsequent booking within a set number of days after the focal transaction. Columns (1), (3), and (5) show estimates for the extensive margin and fail to find economically or statistically meaningful effects. Columns (2), (4), and (6) display results for the intensive margin. There are larger in percentage terms and statistically significant effects for the 2 month and 4 month horizon. At the 12 month horizon, results are similar in levels but standard errors are much wider.

The treatment may also have affected the behavior of hosts. We measure whether hosts change their listing page in response to review related information. Specifically, we measure whether the number of photos or the length of a listing's description changed due to the treatment. Table C.6 shows precisely estimated null effects, meaning that, at least in terms of how hosts advertise their listing, there is no effect.

B.4 Why Do Reviews Affect Views?

In this section, we investigate the mechanisms behind the fact that the treatment group has more views than the control group. There are two main hypotheses for why this effect exists. The first is that searchers can see the number of reviews on the search page, and are induced to click on the listing because of this information. The second is that the ranking algorithm may take into account reviews and display reviewed listings higher.

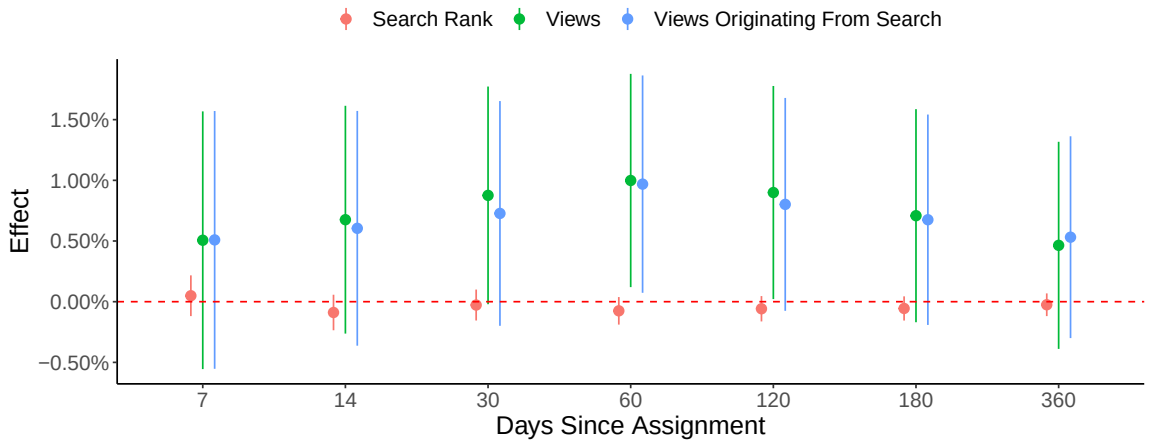
To disentangle this, we measure whether a view originated from search and the search ranking of the listing on a search page prior to a click onto a listings. Figure B.2 shows the effects of the treatment on views originating from a search. These effect on views from search is similar to the effects on overall views. We also measure the effect on the originating search rank. We find a precise zero effect on the search ranking of listings prior to a view.

These results show that views to a listing increase in the treatment but the search ranking does not change. We conclude that the presence of information about reviews in search results matters. Searchers see that treated listings have more reviews and this induces them to click on their their listing page to view more information.

B.5 Home Improvement Platform Scrape

In 2018, Farronato et al. (2020) performed a comprehensive web-crawl of a large home improvement services platform. They identified the largest three cities for each state in terms of unique

Figure B.2: Effect on Search Rank and Views from Search



Notes: This figure plots observational estimates and 95% confidence intervals of the effect of the incentivized review email on views of a listing's page, views originating from search, and the search rank from which the views arrived.

home improvement professionals in categories subject to licensing, and joined that list with the top 100 cities in terms of overall platform activity as measured by the number of requests. Cities with fewer than 10 professionals were excluded. For each category and city, they found the corresponding landing page for the platform. They then obtained information about all professionals displayed on the landing page and their reviews.

We use this crawled dataset to measure the speed of reviews. In a sample of 35,829 professionals, we find that the median time between the first and second review is 10 days. Similar to the result of 6 days for Airbnb. We also find that of those professionals who have one review, 89% have a second review. This demonstrates that sellers who can obtain one transaction can typically obtain additional transactions and reviews, and that these come soon after the first review.

B.6 Heterogeneity by Review Rating

Next, we investigate whether heterogeneous effects are due to some listings receiving good reviews and other listings receiving bad reviews can explain our results. Note that we cannot take an approach similar to the one above, since it is difficult to predict ratings and since submitted ratings are endogenous. Instead, we turn to a calibration exercise. We know from section 4 that the treatment increased the likelihood of a review with rating, r, by an amount $z(r)$. If we also knew the causal effect of a review with rating r, $\tau(r)$ relative to no review on an outcome, Y, then we could calculate the intent to treat effect using the following equation:

$$E[Y|T=1] - E[Y|T=0] = \sum_{r \in \{1,2,3,4,5\}} \tau(r) z(r) \tag{10}$$

Although we don't know $\tau(r)$, we can use multiples of the observational estimates as a benchmark. In particular, suppose we use a linear regression to predict future demand as a function of the star rating, and treat the coefficient on the rating as an estimate of $\tau(r)$. Figure C.10 displays the observational estimates of the effect of a review in the control group on 120 day nights and revenue. We see that five star reviews are associated with much higher demand relative to no review, while one, two, and three star reviews are associated with much lower demand. Note that these estimates are likely to be biased upward in magnitude even after adding controls, since the rating is correlated with factors observable to guests but not to the econometrician. To account for this, we can also test the sensitivity of our calibration to estimates of $\tau(r)$ which are shrunken towards 0 by a factor $k < 1$.

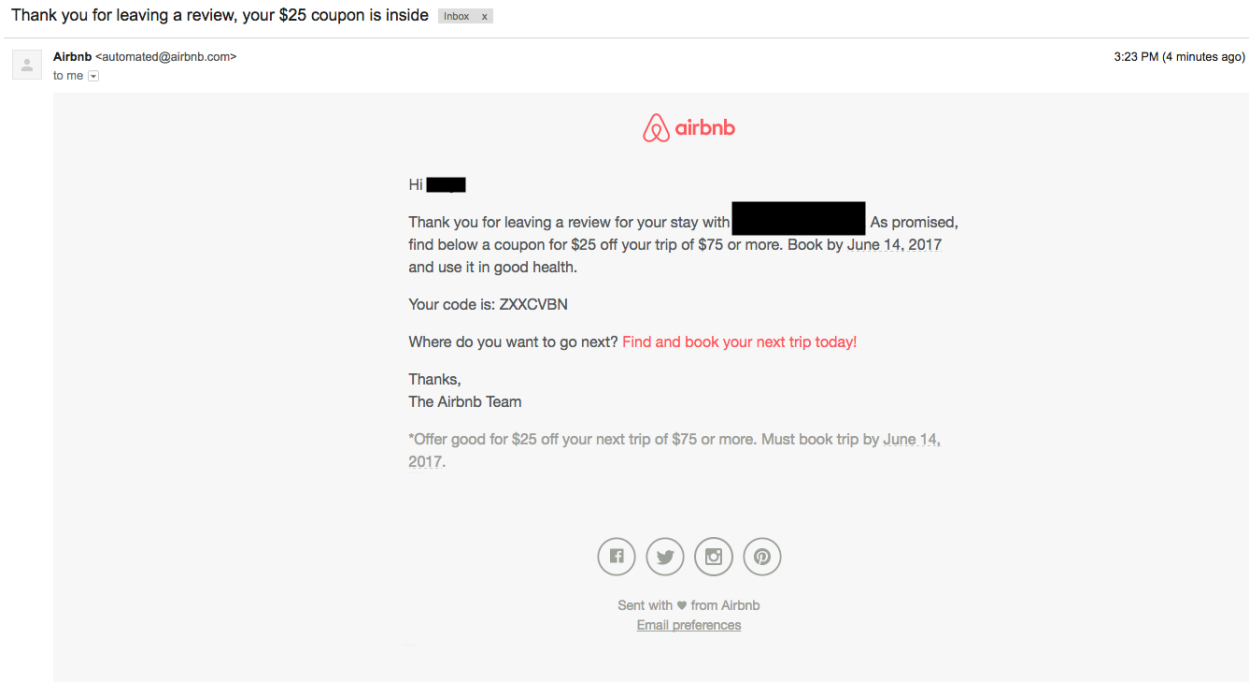
We plug in the observational estimates with controls into Equation 10 and obtain a calibrated estimate of 0.2 for the treatment effect on nights. This estimate is much larger than the regression estimates of 0.02 on nights and is outside of the 95% confidence interval. We then consider shrinkage factors of .5 and .25, for which we find predicted effects on nights of 0.1 and 0.05 respectively, which are still larger than the estimated treatment effects.[14]

Both exercises in this subsection have failed to find that heterogeneity in effects can explain the small and statistically insignificant intent to treat effects on nights and revenue. As a result, we conclude that the effects of incentivized first reviews on listing demand are typically small and that naive observational estimates of the effects of reviews are mostly explained by selection bias.

[14] Using shrinkage factors of 1, .5, and .25, we find expected effects on revenue of \$17, \$8 and \$4 respectively. The point estimate of the treatment effect is, in contrast, \$4.26, although it is less precisely estimated.

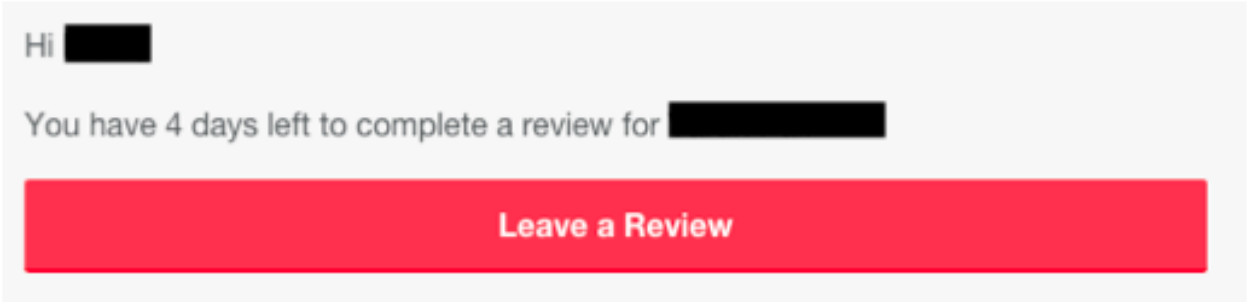
C Appendix: Additional Figures and Tables

Figure C.1: Email Sent After Incentivized Review



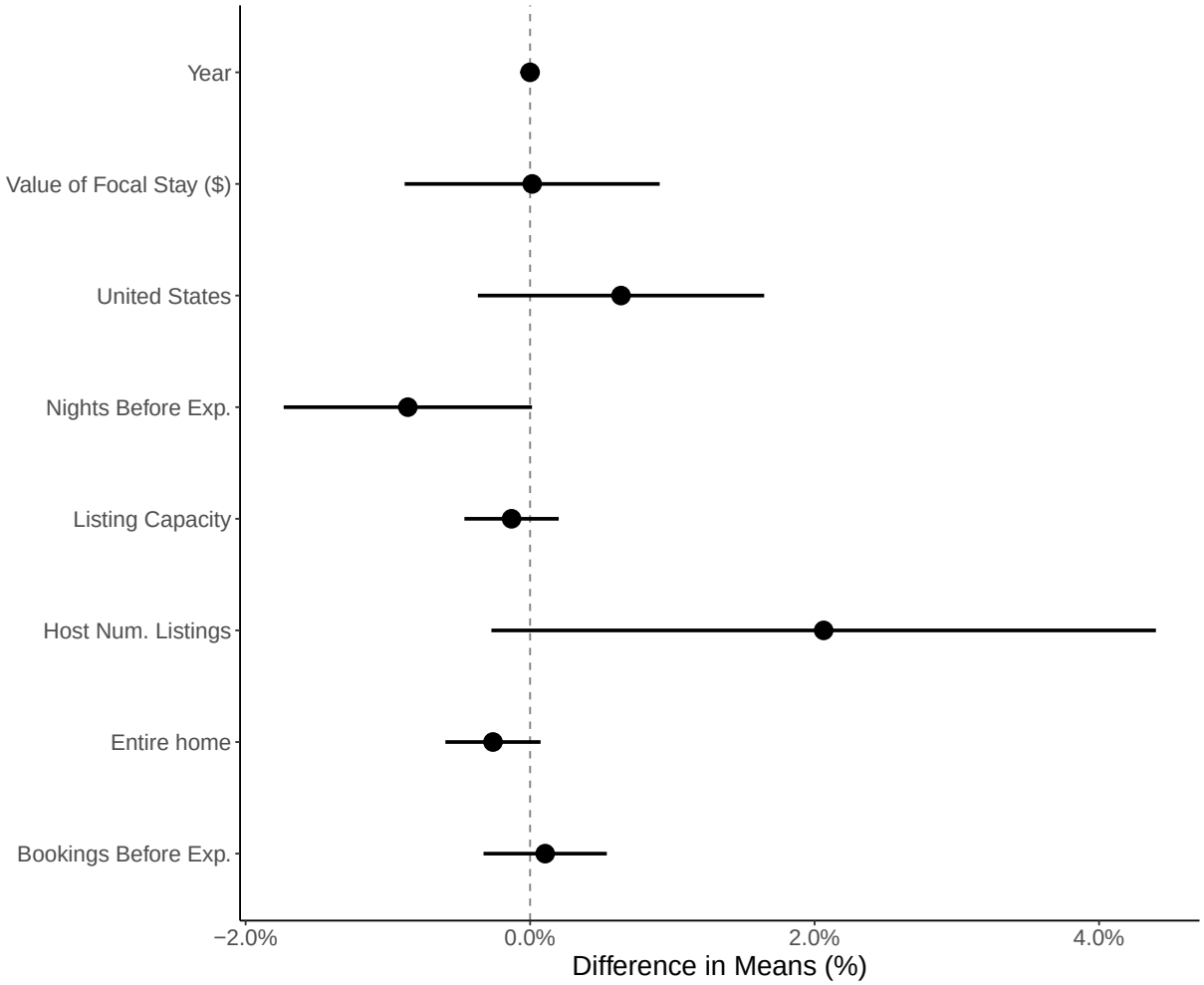
Notes: Displays the email sent to guests who had stayed in treatment listings that had not yet received a review on Airbnb after a certain number of days, issuing them a coupon after leaving a review.

Figure C.2: Email Sent to the Control Group



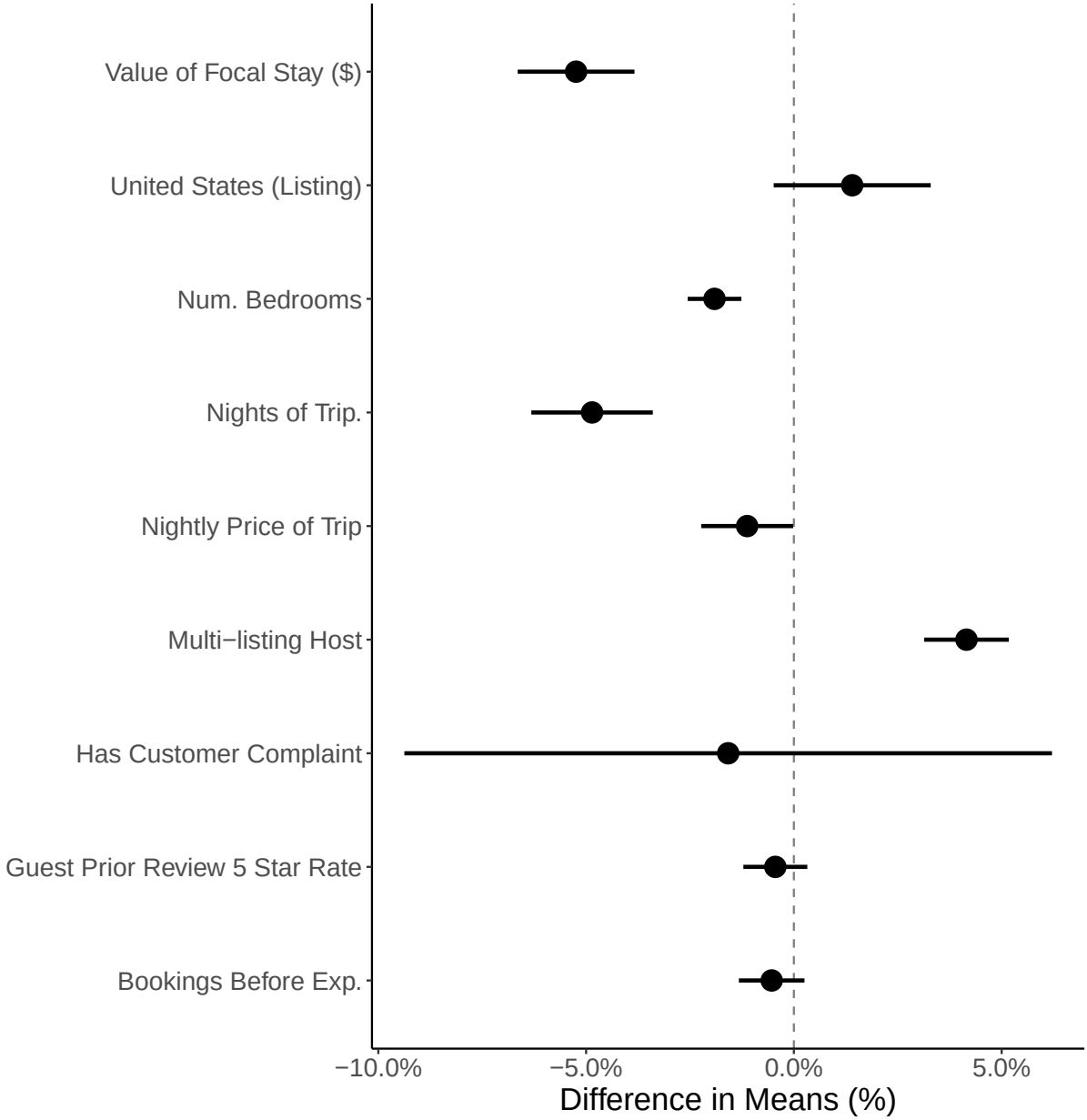
Notes: Displays the email sent to guests who had stayed in control listings that had not yet received a review on Airbnb after a certain number of days.

Figure C.3: Balance Assessment for Experiment



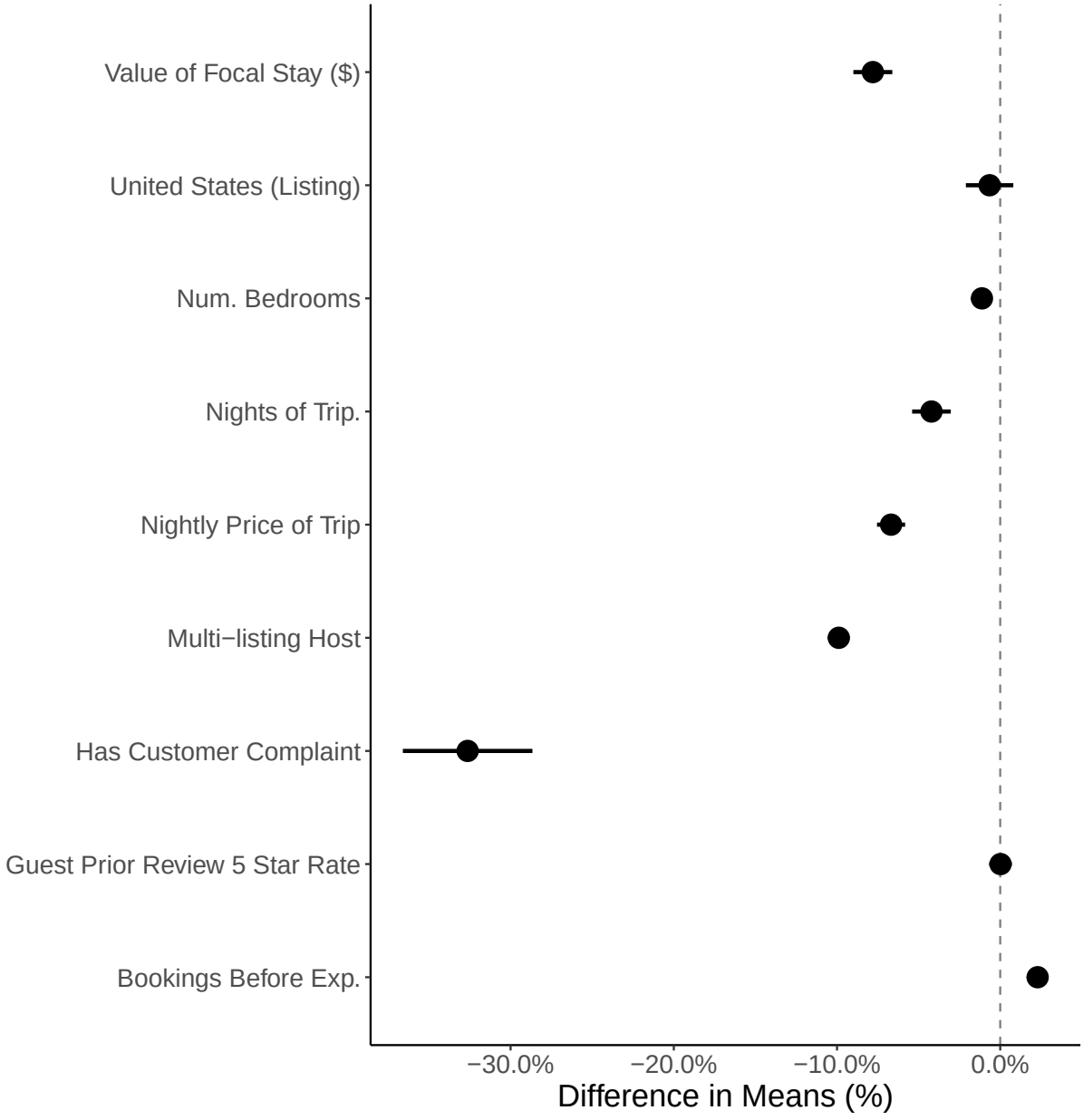
Notes: This plot displays the difference in means between the treatment and control groups for pre-treatment covariates and the days between checkout and email date. No differences were statistically significant in the pre-treatment covariates (year of stay, dollar value of transaction, whether the listing was in the United States, the number of nights the listing hosted for prior to the experimental assignment, the listing capacity, the number of listings by the host, whether the listing was an entire property and the number of bookings prior to the experiment). The days between (coupon / reminder) email and checkout is measured for a subset of listings and exhibits a slight and statistically significant difference between treatment and control.

Figure C.4: Differences in Characteristics of Reviewed Transactions
Treatment vs Control



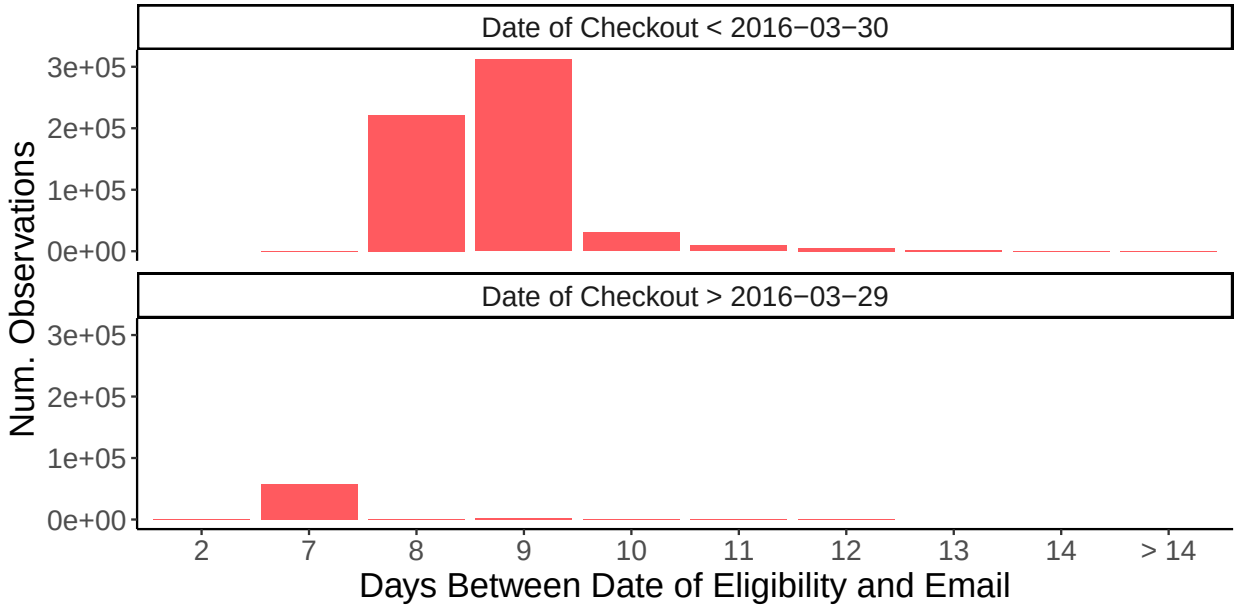
Notes: This plot displays the difference in means between the treatment and control groups for trip characteristics.

Figure C.5: Differences in Characteristics of Transactions
Reviewed vs Non-Reviewed



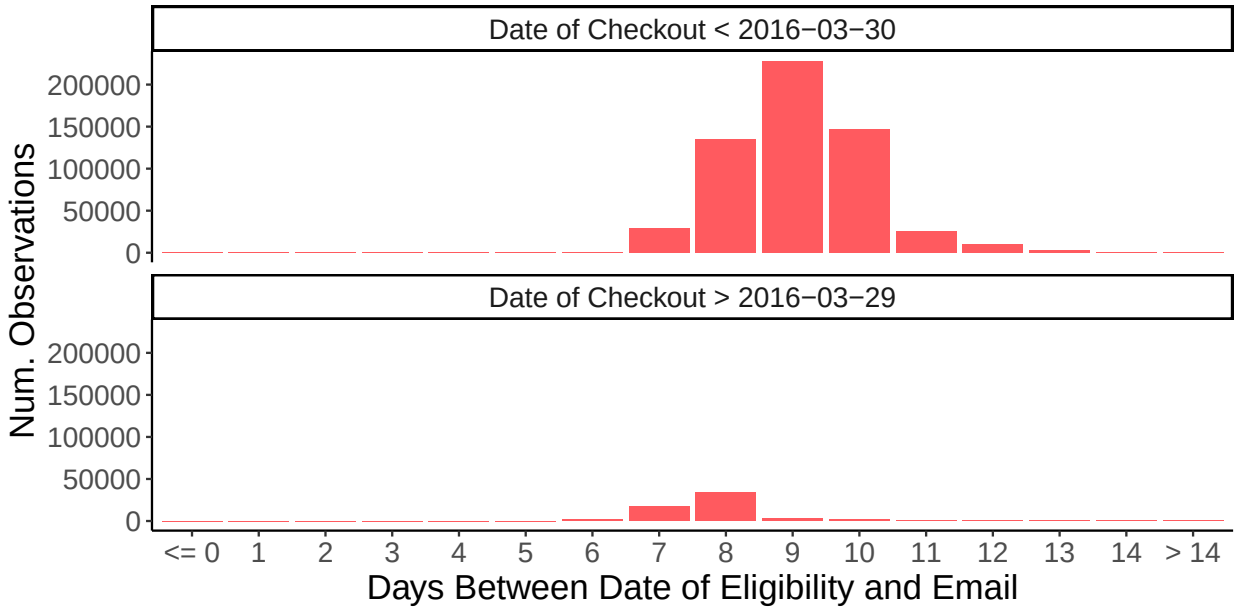
Notes: This plot displays the difference in means between reviewed and non-reviewed transactions in the treatment versus the control groups.

Figure C.6: Days Between Assigned Date and Email



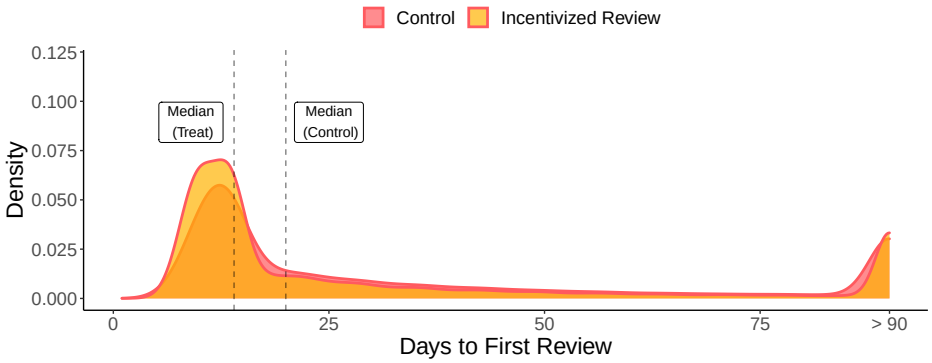
Notes: This figure plots the histogram of days between email and the assigned checkout used by the email dispatch system. Note that no email was logged for 2.2% of observations, either due to missing logging or email dispatch errors.

Figure C.7: Days Between Realized Checkout and Email



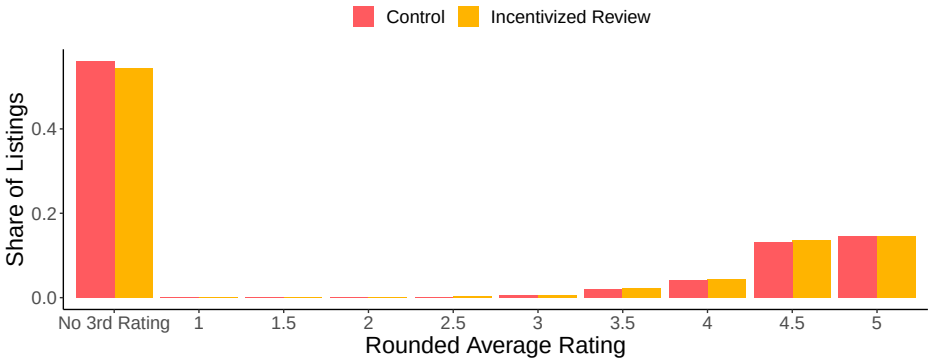
Notes: This figure plots the histogram of days between email and the realized checkout of the focal transaction. Note that no email was logged for 2.2% of observations, either due to missing logging or email dispatch errors.

Figure C.8: Distribution of Days to First Review



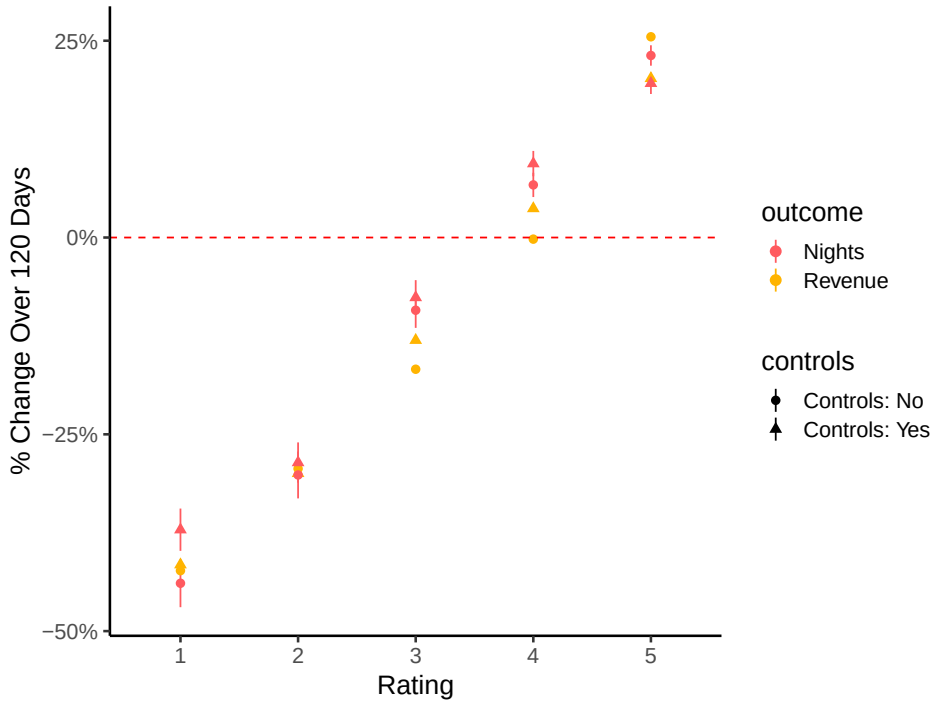
Notes: The figure plots the distribution of the time of arrival for the first review for treated and control listings. The time is calculated as the difference in days between the date of the arrival of the first review and the checkout of the transaction for which the experimental assignment occurred.

Figure C.9: Rounded Average Rating After Three Ratings



Notes: The figure plots the distribution of the rounded average of the first three ratings from any transaction in the treatment and control groups.

Figure C.10: Observational Estimate of Effect of Review



Notes: This figure plots observational estimates and 95% confidence intervals of the effect of a first review with a given star rating (1 - 5) on subsequent nights and revenue. Estimates without controls are represented by circles while estimates with controls are represented by triangles. Controls for room type, capacity, bedrooms, prior nights, prior bookings, trips in process, number of listings managed by the host, main photo size, number of photos, guest gender, whether the guest is a host, guest prior nights, guest prior reviews submitted, guest prior five star reviews submitted are included along with checkout week and market fixed effects.

Table C.1: Effects of Treatment on Demand with Covariates

	(1) Views	(2) Reservations	(3) Nights	(4) Booking Value
Assigned to Treat.	8.876**	0.0431**	0.0182	1.614
	(3.477)	(0.0173)	(0.0689)	(9.450)
R^2	0.31250	0.34640	0.30278	0.32735
Observations	649,266	649,266	649,266	649,266
Controls	✓	✓	✓	✓
Checkout Week FE	✓	✓	✓	✓
Zip Code FE	✓	✓	✓	✓

Notes: This table displays linear regression estimates measuring the effects of the treatment (the guest receiving an email with an offer of a coupon in exchange for a review) on measures of demand. 'Listing Views' refers the number of times the listing's page was viewed, 'Reservations' refers to the number of transactions, 'Nights' refers to the number of nights that the listing was occupied, and 'Booking Value' is the amount paid by guests for transactions involving this listing. All four metrics are calculated for outcomes up to 120 days since the assignment end of the focal transaction. The focal transaction is the first transaction for a listing for which it was eligible for the experiment. Controls for room type, capacity, bedrooms, prior nights, prior bookings, trips in process, number of listings managed by the host, main photo size, number of photos, guest gender, whether the guest is a host, guest prior nights, guest prior reviews submitted, guest prior five star reviews submitted are included along with checkout week and zip code fixed effects.

Table C.2: Effects of Treatment on Trip Characteristics

	Nights Per Trip (1)	Trip Revenue (2)	Price Per Night (3)	Lead Time (Days) (4)
(Intercept)	4.207***	396.5***	103.7***	17.29***
	(0.0098)	(1.340)	(0.3128)	(0.0384)
Treatment	-0.0403**	-3.473	-0.4688	0.0272
	(0.0136)	(1.882)	(0.4403)	(0.0543)
R^2	7.84×10^{-6}	7.37×10^{-6}	4.68×10^{-6}	4.79×10^{-7}
Observations	2,389,288	2,389,288	2,389,288	1,892,755

Notes: This table displays regressions at a transaction level of transaction characteristics on the treatment. All transactions for listings in the experiment that occur within 120 days of the checkout of the focal stay are considered. The regression for lead time includes fewer observations since we considered only trips for which the checkin occurred within 120 days. Standard errors are clustered at the listing level.

Table C.3: Intensive Margin Regression

	Has Res. (1)	Num. Res. (2)	Has Res. (3)	Num. Res. (4)	Has Res. (5)	Num. Res. (6)
Constant	0.5081***	4.191***	0.5846***	6.270***	0.7040***	12.34***
	(0.0009)	(0.0121)	(0.0009)	(0.0186)	(0.0008)	(0.0383)
Assigned to Treatment	0.0016	0.0341**	0.0012	0.0586**	0.0006	0.0621
	(0.0012)	(0.0171)	(0.0012)	(0.0262)	(0.0011)	(0.0541)
Observations	654,595	333,125	654,595	383,029	654,595	461,008

Notes: This table displays OLS regression estimates measuring the effects of being assigned to treatment on intensive and extensive margin outcomes. 'Has Res.' is a binary indicator for whether the listing has received a reservation after being assigned to the experiment and within a given time period (60, 120, and 360 days respectively). 'Num. Res.' is the number of reservations after being assigned to the experiment, for the subsample of observations that have at least one reservation in the time period after the experiment assignment. Robust standard errors are reported.

Table C.4: Effects of Treatment on Transaction Quality - With Covariates

	Complaint (1)	Reviewed (2)	Star Rating (3)
Treatment	-4.26×10^{-5}	0.0047***	-0.0050**
	(0.0001)	(0.0008)	(0.0019)
R^2	0.00373	0.03499	0.02976
Observations	2,431,085	2,431,085	1,579,132
Controls	Yes	Yes	Yes
Guest Region FE	✓	✓	✓
Checkout Week FE	✓	✓	✓
Num. Nights FE	✓	✓	✓
Num. Guests FE	✓	✓	✓

Notes: This table displays regressions measuring the effects of the treatment (the guest receiving an email with an offer of a coupon in exchange for a review) on measures of transaction quality. The set of transactions considered for this regression includes all transactions for which the checkout date was between the checkout date of the focal transaction and 360 days after. 'Complaint' refers to whether a guest submitted a customer service complaint to Airbnb, 'Reviewed' refers to whether the guest submitted a review, 'Star Rating' refers to the star rating of any submitted reviews. Control variables include the log of transaction amount, the number of times the guest has reviewed and reviewed with a five star ratings in the past, the prior nights of the guest, whether the guest has an about description, and guest age on the platform

Table C.5: Heterogeneity Analysis - By Covariate

	Reservations Within 120 Days					
	(1)	(2)	(3)	(4)	(5)	(6)
(Intercept)	2.315***	3.675***	3.668***	3.569***	6.216***	2.097***
	(0.0105)	(0.0124)	(0.0183)	(0.0237)	(0.0614)	(0.0110)
Treatment	0.0356*	0.0373*	0.1269	0.0376*	0.0397*	0.0359*
	(0.0169)	(0.0175)	(0.4053)	(0.0175)	(0.0175)	(0.0169)
Age $<$ 30 Days	3.592***					
	(0.0280)					
Treatment $\times$ Age $<$ 30 Days (Demeaned)	0.0491					
	(0.0396)					
Superhost		2.581***				
		(0.1382)				
Treatment $\times$ Superhost (Demeaned)		0.0523				
		(0.1922)				
Multi-listing Host			0.0842***			
			(0.0248)			
Treatment $\times$ Multi-listing Host (Demeaned)			0.0077			
			(0.0351)			
Female Host				0.0481		
				(0.0303)		
Male Host				0.3783***		
				(0.0326)		
Treatment $\times$ Female Host (Demeaned)				0.0336		
				(0.0428)		
Treatment $\times$ Male Host (Demeaned)				0.0223		
				(0.0461)		
Log Price					-0.5649***	
					(0.0130)	
Treatment $\times$ Log Price (Demeaned)					-0.0427*	
					(0.0185)	
$>$ 1 Booking Prior						3.551***
						(0.0253)
Treatment $\times$ $>$ 1 Booking Prior						0.0330
						(0.0357)
Observations	640,893	640,786	640,936	640,854	640,936	640,936
R^2	0.06324	0.00223	4.59×10^{-5}	0.00058	0.00492	0.06416

Notes: This table displays estimates of heterogenous treatment effects on reservation within 120 days of the focal stay. 'Treatment' refers to the guest of the focal transaction being sent an email offering a coupon. 'Age $<$ 30 Days' refers to a listing being after for fewer than 30 days prior to the focal checkout. 'Multi-listing host' refers to a host having more than 1 active listing. In the gender heterogeneity regressions, the omitted category no gender information. 'Log Price' is the log of the nightly price paid by the guest (inclusive of fees). '$>$ 1 Booking Prior' takes the value of 1 if the listing had more than 1 booking prior to checkout of the focal stay.

Table C.6: Change in Listing Characteristics Over a Year

	Num. Photos Changed	Description Length Changed
	(1)	(2)
(Intercept)	0.3580***	0.4324***
	(0.0008)	(0.0009)
Treatment	-0.0006	0.0011
	(0.0012)	(0.0012)
Observations	653,907	653,907

Notes: This table the results of a linear regression where the outcome variable is whether the number of photos or the length of the description changed for listings between the start of the treatment and 360 days later. Fewer than 1000 observations were dropped because they could not be matched with listing photos and descriptions in the database.