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

Andrey Fradkin* David Holtz‡

August 25, 2022

Abstract

Many online reputation systems operate by asking volunteers to write reviews for free. As a result, a large share of buyers do not review, and those who do review are self-selected. This can cause the reputation system to miss important information about seller quality. We study the extent to which a platform can improve market outcomes by attempting to increase the amount and quality of information collected by its reputation system. We do so by analyzing a randomized experiment conducted by Airbnb. In the treatment, buyers were offered a coupon to review listings that had no prior reviews. In the control, buyers were not offered any incentive to review. We find that although the treatment induced additional reviews that were more negative on average, these reviews did not affect the number of nights sold or total revenue. Furthermore, we find that, contrary to the treatment's intended effect, Airbnb's incentivized program caused transaction quality for treated sellers to fall. We examine how the quality of the induced reviews, market conditions, and the design of Airbnb's reputation system can explain our findings.

*fradkin@bu.edu, Corresponding Author

†dholtz@haas.berkeley.edu

‡We thank Dean Eckles, Chiara Farronato, Shane Greenstein, John Horton, Caroline Hoxby, Xiang Hui, Ramesh Johari, Garrett Johnson, Tobias Klein, Jon Levin, Tesary Lin, Mike Luca, Steve Tadelis, Catherine Tucker, Giorgos Zervas, Olivia Natan, 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. While existing reputation systems are considered critical to the success of digital marketplaces, they suffer from a variety of imperfections documented in previous research (Tadelis, 2016). A major imperfection present in many reputation systems is that because online reputation is a public good,[1] not all buyers leave reviews, and the subset who do are self-selected.

In theory, increasing review rates can improve market outcomes for at least two reasons (Avery, Resnick and Zeckhauser, 1999). First, the more often reviews are left by buyers, the faster other buyers learn about sellers' quality. Acemoglu et al. (2022) show that increasing the speed of learning about seller quality increases welfare by facilitating better matches. Second, when the reviews collected by a reputation system are non-representative, buyers may form biased beliefs about the quality of sellers, leading to the formation of bad matches. For instance, if reviews are more likely to be submitted after a positive experience than a negative one, future buyers may overestimate the quality of sellers (Dellarocas and Wood, 2007; Nosko and Tadelis, 2015). In concordance with this theory, platforms such as Amazon, Wayfair, and Walmart use incentives to improve review rates.[2] However, the effectiveness of these incentivized review programs in helping the market is unknown.

Using a large scale and costly experiment conducted on Airbnb, we study whether incentivizing reviews improves market outcomes and find that it does not.[3] In the treatment group, guests to treated listings were sent an email offering them a $25 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 incentivized review policy increased the review rate by 53% relative to the review rate in the control group, and induced reviews with lower star ratings on average. These results con-

[1] Each individual review helps other market participants, but not the person who provides the review.

[2] As an example, Figure C.1 shows an email sent as part of an incentivized review program run by the eCommerce website Girlfriend Collective.

[3] We were not able to calculate the exact cost of the experiment since data on coupon redemption was not available to us, but our best guess is that the policy cost on the order of $250,000.

firm prior work that shows that incentives and nudges are effective at inducing additional reviews (Burtch et al., 2018; Marinescu et al., 2021; Karaman, 2020), and that reviews induced by incentives are more likely to contain information about lower quality transactions (Burtch et al., 2018; Marinescu et al., 2021). We also find that compared to non-incentivized reviews with the same star rating, incentivized reviews corresponded to worse guest experiences. In other words, although incentivized reviews had a lower average rating, they were more inflated at a given star rating than non-incentivized reviews.

Although the incentivized review program impacted the number and type of reviews left by guests, it had neutral to negative impacts on the business outcomes of the platform and its sellers. Incentivized reviews did not affect the total quantity (nights) sold for treated listings, but did cause a change in the composition of transactions. Specifically, treated listings had more transactions, but stays at treated listings lasted fewer nights on average. As a consequence, the revenue effects of incentivized reviews are not statistically distinguishable from zero. Furthermore, contrary to the policy's intent, incentivized reviews 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. We argue that this is, in part, related to our finding that incentivized reviews were less correlated with true transaction quality.

Our inability to detect any positive seller- or market-level effects stemming from the introduction of the incentivized review policy can be attributed both to market structure and to the design of Airbnb's reputation system. On the market structure front, 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 unless they receive an initial review, then the effect of incentivized reviews will be large. We show that reviews from other transactions arrived quickly for listings in our experimental sample. Furthermore, listings on Airbnb have capacity constraints that limit the extent to which any treatment can increase quantity sold. While our empirical results are specific to Airbnb, fast review arrival rates and

capacity constraints are common to many large online marketplaces.[4]

Incentivized reviews may have also had 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 star and were only displayed once a listing had at least three ratings. As a result, ratings induced by the incentivized review program were averaged with at least two other ratings when displayed to guests on Airbnb. This averaging, rounding, and censoring attenuated perceived differences between the ratings of control and treatment listings. We expect similar mechanisms to exist in other platforms which round to half stars, including Amazon, Etsy, and Yelp.

An important implication of our findings is that institutional details such as market conditions and reputation system design are critical for understanding the value of reviews and the effects of reputation system-related treatment interventions. We do not claim that that reviews and reputation systems have little value. In fact, prior work has shown that reputation systems substantially increase consumer surplus (Wu et al., 2015; Lewis and Zervas, 2016; 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 rounded averages.

2 Literature Review

We contribute to several related research themes within the study of online reputation systems. The first and most directly relevant of these research themes focuses on incentivized review policies. In an early contribution to the literature studying online reviews, Avery, Resnick and Zeckhauser (1999) show that, in equilibrium, evaluations will be underprovided relative to what is optimal from a welfare perspective unless would-be reviewers are offered an incentive to leave feedback. Building on this insight, 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.

[4]We document a similarly fast review arrival rate on a large online marketplace for home improvement services in Appendix B.8.

(2018), Marinescu et al. (2021), and Karaman (2020) experimentally document that incentives and solicitations offered by platforms can not only increase review rates, but also generate more representative reviews.[5] In an adjacent stream of work, Li (2010), Li and Xiao (2014), Cabral and Li (2015), and Li, Tadelis and Zhou (2020) study policies in which sellers (rather than the platform) offer rebates for reviews.

In a piece of research that is particularly closely related to our work, Pallais (2014) experimentally measures the effects of an intervention in which new sellers in an online labor market 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. An advantage of this policy relative to the one studied in Pallais (2014) is that it is less expensive for platforms to implement since they do not have to pay to hire new sellers. However, a disadvantage is that the incentivized first review policy does not solve the 'cold-start' problem for listings that are unable to transact prior to receiving their first review.

Beyond work that studies incentivized review policies and nudges to review, there is a large body of work that that studies bias in reputation systems. Several papers show that non-incentivized reviews exhibit a positivity bias (Dellarocas and Wood, 2007; Nosko and Tadelis, 2015; Filippas, Horton and Golden, 2022; Brandes, Godes and Mayzlin, 2019; Fradkin, Grewal and Holtz, 2021). This positivity bias has a number of causes, including a higher propensity to post feedback when satisfied (Dellarocas and Wood, 2007), reciprocity and retaliation (Dellarocas and Wood, 2007; Fradkin, Grewal and Holtz, 2021), selection effects that dictate which buyers transact with which sellers (Nosko and Tadelis, 2015), 'reputation inflation' (Filippas, Horton and Golden, 2022), and differential attrition of those with moderate experiences (Brandes, Godes and Mayzlin, 2019).[6] Our results highlight that positivity bias in incentivized ratings may even be worse than in non-incentivized ratings. In particular, conditional on a given star rating, reviews induced by the incen-

[5] Other work that uses field experiments to study the effects of changes to reputation system design includes (Cabral and Li, 2015; Benson, Sojourner and Umyarov, 2020; Fradkin, Grewal and Holtz, 2021; Hui, Liu and Zhang, 2020; Garg and Johari, 2021). A preliminary analysis of this experiment focusing on the first month was presented in Fradkin et al. (2015).

tive tended to correspond to worse experiences.

There is also an emerging research literature that aims to model the dynamics with which consumers learn about seller quality under different ratings systems (Papanastasiou, Bimpikis and Savva, 2018; Besbes and Scarsini, 2018; Acemoglu et al., 2022; Ifrach et al., 2019). Our work is closely related to Besbes and Scarsini (2018) and Acemoglu et al. (2022), which both compare consumer learning dynamics under different reputation system designs. We will argue in this paper that one of reasons that the additional reviews induced by Airbnb's incentivized first review program failed to have a measurable impact on market outcomes is the design of Airbnb's review system, which has characteristics of the "summary statistics only" systems discussed in both Acemoglu et al. (2022) and Besbes and Scarsini (2018).[7]

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 the number and type of reviews left by guests, but also on subsequent market outcomes.[8] Although one might expect an incentivized first review program like the one we study to generate additional information on seller quality that yields more and better matches on the platform, we do not find this to be the case empirically; we instead find that the program had no effect on demand and negative effects on match quality. We argue that this is due to market conditions in the markets where most of our sample listings were located, the manner in which Airbnb's reputation system aggregates and displays review information, and the capacity constraints of Airbnb listings.

[6]A related literature studies 'fake' reviews (Luca and Zervas, 2016; Mayzlin, Dover and Chevalier, 2014; He, Hollenbeck and Proserpio, 2020), however, fake reviews are less of a concern in our research setting because all Airbnb reviews are linked to verified transactions.

[7]In Appendix A, we develop a theoretical framework that simplifies the framework put forth in Acemoglu et al. (2022) and clarifies the conditions under which we would expect incentivized reviews to increase demand and/or the utility of buyers.

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

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 the host and the guest 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 listing had received at least three reviews; before that, only review text was visible on a listing page. The number of reviews a listing had received was visible on both the search and listing pages as long as the listing had one review, meaning that reviews were able to have effects both through the search page and through the listing page.

Prior to July 2014, 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 put in place (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 threshold was typically 8 or 9 days throughout most of our sample, with the

specific number of days varying due to idiosyncratic details of Airbnb's email dispatching system.[9]

Across Airbnb's entire platform, guests who had not reviewed within the threshold number of days received an email reminding them to review. For stays at control listings that met the criteria above, guests received the standard review reminder email. For stays at treatment listings that met the criteria above, 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, 326,602 listings were assigned to the control, whereas 328,266 listings were assigned to the treatment. The experiment used a well-tested system at Airbnb for randomization into treatment conditions and achieved good balance on pre-treatment covariates (Figure C.4).

Because randomization was conducted at the listing-level, many of our analyses will utilize the concept of a listing's *focal stay*, which is the first transaction for which a listing is either in the treatment or control. This is in contrast to subsequent stays that also may have resulted in reviews, but may have been affected by the presence of a review for the focal stay.

4 Effects of Experiment on Reviews

Incentivized reviews can only have effects on market outcomes if they have an effect on the quantity and types of reviews that are submitted by guests. Thus, we first measure the effects of incentives on these review-related outcomes. We find that the treatment induced additional reviews for focal stays and that those reviews tended to have lower ratings and more negative text sentiment on average.

Table 1a shows the effect of the treatment on review rates and the distribution of numerical ratings before conditioning a review being left. The first thing that is apparent is that the treatment

[9] More detailed information on the timing of review email dispatch during our experiment is found in Appendix B.1.

Table 1: Differences in Ratings by Treatment

(a) All Trips

	No Review	1 Star	2 Star	3 Star	4 Star	5 Star
	(1)	(2)	(3)	(4)	(5)	(6)
Constant	0.7581***	0.0042***	0.0045***	0.0140***	0.0671***	0.1521***
	(0.0007)	(0.0001)	(0.0001)	(0.0002)	(0.0004)	(0.0006)
Incentivized Review	-0.1286***	0.0020***	0.0031***	0.0144***	0.0464***	0.0628***
	(0.0011)	(0.0002)	(0.0002)	(0.0004)	(0.0007)	(0.0010)
R^2	0.01946	0.00019	0.00040	0.00250	0.00655	0.00657
Observations	654,868	654,868	654,868	654,868	654,868	654,868

(b) Reviewed Trips

	1 Star	2 Star	3 Star	4 Star	5 Star	Positive Text
	(1)	(2)	(3)	(4)	(5)	(6)
Constant	0.0175***	0.0186***	0.0578***	0.2773***	0.6288***	0.9405***
	(0.0005)	(0.0005)	(0.0008)	(0.0016)	(0.0017)	(0.0010)
Incentivized Review	-0.0008	0.0019**	0.0188***	0.0290***	-0.0489***	-0.0080***
	(0.0006)	(0.0006)	(0.0011)	(0.0021)	(0.0022)	(0.0013)
R^2	8.93×10^{-6}	4.25×10^{-5}	0.00131	0.00097	0.00238	0.00026
Observations	200,611	200,611	200,611	200,611	200,611	135,670

Notes: The above tables display the results of regressions in which the outcome is a rating outcome and the only covariate is the treatment indicator. Panel a) displays the regression for all trips while panel b) displays results for just reviewed trips. Note that panel b) column 6) has fewer observations since reviews in a foreign language could not be classified by the sentiment analyzer. Robust standard errors are displayed in parentheses.

was effective at increasing the rate at which guests submitted reviews: the treatment increased the review rate by 12.86 percentage points, from 24.19% to 37.05%. Because of this increase in review rate, before conditioning on a review being left the treatment also increased the number of five-star reviews (6.28 pp), four-star reviews (4.64 pp), three-star reviews (1.44 pp), two-star reviews (0.31 pp), and one-star reviews (0.2 pp).

In addition to increasing the absolute number of reviews left at each star rating, the treatment also changed the relative frequency of ratings. Table 1b shows that conditional on a review, ratings of treated listings were in fact lower than ratings of control listings; the treatment caused the average rating left by guests to drop by 0.07 stars, from 4.48 to 4.41. This effect was driven by a decrease in the rate of five-star reviews and an increase in the rate of two- to four-star reviews. In other words, although the treatment led to an across-the-board increase in the number of reviews at all star ratings, the increase was larger for lower ratings than for higher ratings.

The reviews induced by the incentivized review treatment not only had lower star ratings, they also had more negative text. After classifying the sentiment of the review text for each review in our sample, we find that the percentage of review text classified as positive decreased by 0.8% in the treatment group.[10] This difference in text sentiment disappears once we condition on the review's star rating, suggesting that the effects we report on star ratings and review text are consistent with each other.

The differences in ratings that we observe should correspond to differences in the observable and/or unobservable characteristics of reviewed transactions across the experiment's treatment conditions. Table 2 reports the effect of the incentivized review treatment on the prevalence of different trip- and user-level characteristics among reviewed focal transactions. We find that reviews in the treatment group were more likely to come from trips to lower capacity Airbnb listings (column 5) that were run by multi-listing hosts (column 1). These trips were also on average lower value (column 2), due both to shorter duration (column 3) and lower price per night (column 4). Overall, these effects are consistent with the notion that the subset of guests who were responsive to a $25

[10]We describe our methodology for text classification and the details of our results about sentiment in Appendix subsection B.2.

coupon incentive were more likely to be budget-conscious travelers.

Table 2: Differences in Characteristics of Reviewed Transactions
Treatment vs Control

	Pro. Host (1)	Revenue (2)	Nights (3)	Price (Nightly) (4)	Bedrooms (5)	Complaint (6)	GFSR (7)
Constant	0.4499***	639.3***	6.738***	116.0***	1.602***	0.0128***	0.6894***
	(0.0018)	(3.681)	(0.0400)	(0.5065)	(0.0042)	(0.0004)	(0.0021)
Incentivized	0.0187***	-33.52***	-0.3274***	-1.308*	-0.0307***	-0.0002	-0.0031
	(0.0023)	(4.730)	(0.0514)	(0.6507)	(0.0053)	(0.0005)	(0.0027)
R^2	0.00034	0.00025	0.00020	2.02×10^{-5}	0.00016	7.8×10^{-7}	1.56×10^{-5}
Observations	199,654	199,654	199,654	199,654	199,654	199,654	82,182

Notes: The above tables display the results of regressions in which the outcome is a trip characteristic and the regressor is whether the review was incentivized (vs control). 'Pro. Host' is a host with more than one listing, 'Complaint' takes the value of 1 if the customer had a complaint to Airbnb during the trip, and 'GFSR' is the guest's rate of giving five-star ratings for any prior trips they've taken (note that this has fewer observations since many guests had not previously had an Airbnb trip). Robust standard errors are displayed in parentheses.

The lower average ratings we observe in induced reviews could also be driven, as posited by much of the literature, by lower quality transactions. We can't directly observe transaction quality, but we do have a sparse proxy for very bad transactions — the customer complaint rate. In column 6 we show that customer complaint rates did not differ in a statistically significant manner between treatment and control transactions. Note that this lack of difference does not preclude differences in transaction quality that may not be captured by customer complaints.

One other reason that incentivized reviews may have different ratings than non-incentivized reviews is that the guests who review due to incentives may have had different reviewing styles. For example, these guests may tend to judge Airbnb listings more harshly than other reviewers. We test for this possibility by measuring the reviewing behavior of guests prior to the focal transaction. Specifically, we comparing the average historical guest-level five-star review rate (GFSR) for reviewed trips in the control and treatment groups. We find that the observed difference in review ratings and text is not explained by guest harshness (column 7).

Although the incentivized review treatment reduced the amount of differential non-response in reviewing, it did not eliminate it entirely. This is apparent in Figure C.7, which compares the trip- and user-level characteristics of reviewed transactions in the treatment group to those of non-

reviewed transactions in the treatment group. First, reviewed transactions in the treatment were less likely to have complaints than non-reviewed transactions in the treatment, implying that even under the incentivized review policy many extremely low quality transactions were not reviewed. Second, reviewed transactions in the treatment had lower value, duration, and price per night than non-reviewed transactions in the treatment, further supporting the notion that less budget-conscious guests are not as responsive to the $25 Airbnb coupon incentive.

In summary, the incentivized review treatment induced additional reviews, and those reviews were on average more negative. These reviews tended to come from guests who were more price sensitive, and thus, more responsive to the offer of a $25 Airbnb coupon. Although the provision of these reviews reduced the severity of the selection bias in which guests leave feedback, there is evidence that even in the treatment, many guests who were less price sensitive and/or who experienced low quality transactions still did not review.

5 Effects of Incentivized Reviews on Market Outcomes

We've shown that the treatment induced reviews, which changed the information set of subsequent buyers. It is, however, theoretically ambiguous if and how these reviews affected market outcomes.[11] In this section, we measure the extent to which the incentivized review program changed outcomes such as purchase rates, platform revenues, and match quality.

5.1 The Effect of Incentivized Reviews on Demand for a Listing

We first measure the effect of the treatment on a listing's subsequent demand. Although prior literature has shown that positive (negative) reviews increase (decrease) demand, the direction and magnitude of the effect of our treatment on demand is ambiguous. This is both because the review incentive induced a mixture of positive and negative reviews, and because the extent to which an individual positive or negative review impacts demand in our setting is unknown. To measure the

[11] In Appendix A, we use a theoretical model to explore the exact circumstances under which reviews induced by the incentives will increase market outcomes.

effects of the treatment on demand, we estimate linear regressions of the following form:

$$y_l = \beta_0 + \beta_1 T_l + \epsilon_l \tag{1}$$

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 is one of the following proxies for guest demand in the 120 days immediately following the focal trip checkout: the total number of listing views, the total number of transactions, the total number of nights across all transactions, and the total booking value across all transactions.

Table 3: Effects on Listing Outcomes (120 Day Horizon)

(a) Intent to Treat

	Views (1)	Reservations (2)	Total Nights (3)	Booking Value (4)
Constant	753.3***	3.665***	15.04***	1,638.6***
	(2.317)	(0.0121)	(0.0469)	(6.525)
Assigned to Treatment	6.725*	0.0416*	0.0227	4.262
	(3.359)	(0.0171)	(0.0661)	(9.202)
Observations	654,595	654,595	654,595	654,595

(b) Local Average Treatment Effect

	Views (1)	Reservations (2)	Total Nights (3)	Booking Value (4)
Constant	740.6***	3.586***	14.99***	1,630.6***
	(8.118)	(0.0418)	(0.1616)	(22.48)
Reviewed	52.70*	0.3259*	0.1782	33.40
	(26.32)	(0.1344)	(0.5180)	(72.10)
Observations	654,595	654,595	654,595	654,595

Notes: The above table display the results of regressions in which outcomes are measured between the checkout of the focal trip and 120 days afterward. Panel a) displays the intent to treat estimates while panel b) displays local average treatment effect estimates from a two-stage least squares regression where the first stage is a regression of whether a focal trip was reviewed on the treatment assignment. Robust standard errors are displayed in parentheses.

Table 3a shows that views (column 1) and reservations increased (column 2). On the other hand, the total nights of stay (column 3) and booking value (column 4) exhibit small effects that are statistically indistinguishable from 0. Both hosts and the platform care most about total nights

booked and the revenue from transactions, meaning that according to the most meaningful metrics, the treatment did not increase demand on average.[12]

A key component of any incentivized review program is imperfect compliance, meaning that not everyone responds to an incentive by reviewing. If compliance in our experiment were sufficiently low, then we may not find any effect of a review simply because too few additional reviews were produced. To better understand what our experiment implies about the effect of a review, we use a two-stage least squares estimator. In particular, we measure the local average treatment effect of an incentivized review using the following second stage equation:

$$y_l = \beta_0 + \beta_1 R_l + \epsilon_l \tag{2}$$

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.[13]

Table 3b displays our estimates of the local average treatment effect of an incentivized review. We find that incentivized reviews generated more attention and transactions for listings. Specifically, the treatment led to a 7.1% increase in views and a 9.1% increase in transactions, which translates to an additional 0.326 transactions per listing, in the 120 days immediately following the focal checkout. The fact that the treatment increased listing views and transactions by similar percentages suggests that the effect of an incentivized review on transactions was primarily driven by an increase in clicks from the search page to the listing page. In subsection B.4, we show that this change in clickthrough rate was due to the fact that the number of reviews a listing had was displayed on the search page, rather than changes in listings' algorithmic search ranking. We also

[12] In Figure C.10 we display the effects of the incentivized review treatment (in percentage terms) on demand outcomes for time horizons other than 120 days. We see similar results across time horizons, with the peak occurring at 120 days. The effects in percentage terms shrink as the horizon expands, which reflects the temporary effects of the treatment. In subsection B.3 we show that the effect on reservations comes from the intensive margin and that the estimates remain similar when adding controls.

[13] This analysis requires 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). Also 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 12.86 percentage points, the causal effect of the coupon email on the review rate.

find that the increase in transactions due to an incentivized review did not translate to an increase in the number of nights booked, and the 95% confidence interval around our point estimate rules out effects of an incentivized review on nights that are larger than 8%. We view this as evidence against large effects of incentivized reviews on demand.

Given that the number of transactions rose but the number of nights did not, it must be the case that incentivized reviews changed the *types* of trips that occur. We investigate this by estimating trip-level regressions on a dataset consisting of all transactions that were booked within 120 days of assignment, with standard errors clustered at the listing level. Table 4 shows the results of these regressions. We find that while there are no statistically significant differences with respect to most trip characteristics, subsequent trips to treated listings had 1% fewer nights per trip. Put differently, although the treatment increased the number of bookings that occurred, the average booking was shorter in duration, such that there was not a statistically significant treatment effect on total nights or booking value.

To summarize, the net effect of Airbnb's incentivized review program on quantities sold and revenue was statistically indistinguishable from 0, despite the fact that treated listings received more listing views and transactions due to the incentivized review program. This is because although the receipt of an incentivized review led to a greater number of transactions, these transactions were shorter on average.

5.2 Why Don't Incentivized Reviews Affect Demand?

There are several reasons why the incentivized review treatment failed to increase demand in the 120 days following treatment assignment, despite increasing the number of bookings listings received.[14] The first of these is the fact that listings in our sample were typically able to generate transactions even without a review. Thus, trips occurring after the focal transaction provided additional opportunities for listings to receive their first review. Reviews from non-focal trips often arrived quickly, and attenuated differences in the distributions of first ratings for treatment and

Table 4: Characteristics of Subsequent Trips

(a) Intent to Treat

	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)
Assigned to 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

(b) Local Average Treatment Effect

	Nights Per Trip (1)	Trip Revenue (2)	Price Per Night (3)	Lead Time (Days) (4)
(Intercept)	4.290***	403.7***	104.7***	17.23***
	(0.0358)	(4.929)	(1.151)	(0.1394)
Reviewed	-0.3133**	-26.99	-3.643	0.2088
	(0.1059)	(14.64)	(3.422)	(0.4161)
R^2	-0.00041	-0.00022	4.68×10^{-5}	0.00013
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. Panel a) displays intent to treat effects while panel b) displays local average treatment effects from a 2SLS regression. Standard errors are clustered at the listing level.

control listings.

More specifically, because the review incentive was only offered to guests after checkout, every listing in the experiment had by definition been able to receive at least one booking without having any reviews. This means that, at least for some guests, the presence of a first review was not pivotal in their choice of listing. One reason that guests take a chance with a non-reviewed listings is that many Airbnb markets are supply-constrained.[15] As a result, guests are shown listings without reviews and sometimes book these listings.

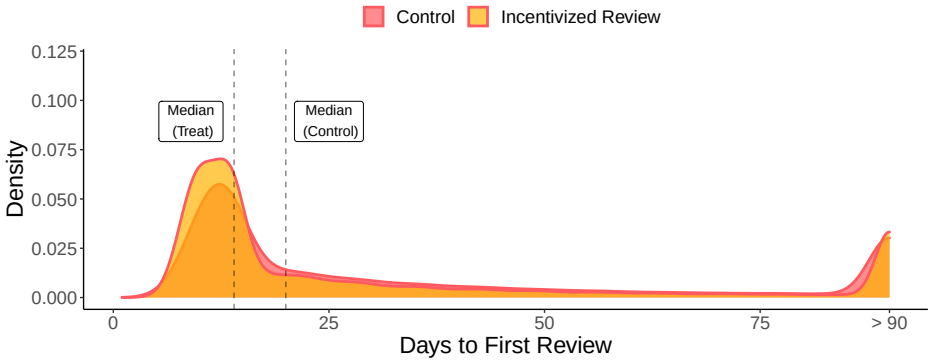
The focal trips that we focus on in our analysis are not anomalies; 45% of listings in the experiment had more than one booking prior to the checkout of the focal trip, not to mention bookings occurring after the focal trip ended. Each of these additional bookings offered an opportunity for the listing to receive a review and these opportunities add up. By August 2016, 72.8% of listings in the control group and 78.4% of listings in the treatment group had received at least one review. This 5.6% difference is less than half as large as the effect of the treatment on the review rate for the focal transaction (13%). Furthermore, not much time elapsed during which treatment listings had a review and control listings did not, because even in the control group, reviews arrived quickly: Figure 1 shows that the difference between the median time to first review in the control and treatment groups is only 6 days.

Even though the effect of the treatment on the percentage of listings receiving at least one review shrinks after taking into account reviews from non-focal transactions, one might still expect differences in the control and treatment first rating distributions to result in demand effects. However, Figure 2 shows that after taking into account reviews from additional bookings, the differences between the first ratings distributions of treatment and control reviews (previously reported in Section 4) are also attenuated. This is likely due to a number of factors, including the fact that reviews from non-focal trips in the treatment consist of a mix of non-incentivized reviews

[14] We also show in Appendices B.6 and B.7 that our lack of overall treatment effect on demand is not obscuring treatment effect heterogeneity with respect to ex-ante expectations of listing demand or the star ratings of the reviews induced by the incentive to review.

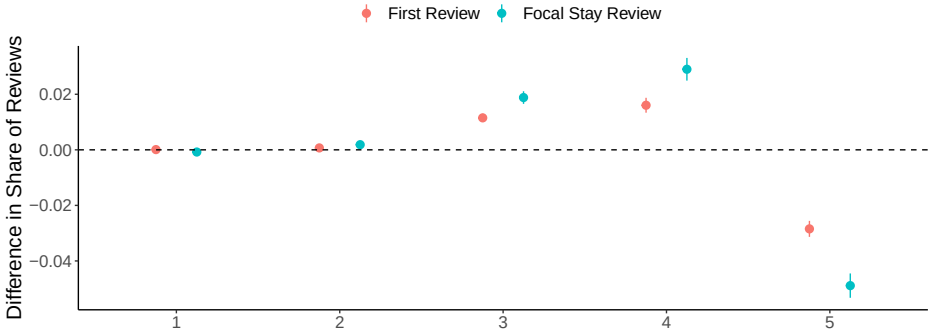
[15] We can measure the degree of supply-constraint by dividing the number of inquiries by the number of listings contacted by market during the time that the experiment was conducted. The average listing in our experiment was booked in a market where the tightness (31.6) was much higher than the tightness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

Figure 1: 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 2: 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 between treatment and control listings. '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. Note that to be eligible for the experiment, the listing must have had no prior reviews.

(left within 8-9 days) and incentivized reviews. Regardless of the reasons behind this attenuation, a smaller difference in the distributions of first ratings that listings received also reduces the likelihood that we would observe an effect of the treatment on demand.

A second reason that incentivized reviews had no effect on demand is the manner in which Airbnb displayed reviews at the time of the experiment. Ratings were not shown for every review, but were instead averaged and rounded to the nearest half star.[16] Furthermore, while review text was publicly visible after a listing received its first review, average ratings were only shown after a listing had received three reviews. Because numerical rating information was not available to some

guests when they viewed the listing page, and for guests who observed numerical information, any differences had been attenuated by at least two other reviews, the likelihood that one review would change a guest's perception of the listing is low.

A third reason is that the capacity of Airbnb listings is limited. In particular, unlike markets for mass-produced goods, only one buyer can book an Airbnb listing per night. As a result, increased listing views caused by the treatment could only increase quantity sold if those listing views resulted in bookings for marginal nights.

In theory, it's also possible that demand was not affected by incentivized reviews because sellers updated their nightly prices after receiving a review. However, our results suggest that Airbnb hosts did not do so.[17] This behavior is consistent with Huang (2021), who finds that sellers on Airbnb are often inattentive or constrained in changing prices when responding to demand fluctuations.

To summarize, a number of explanations contribute to the fact that the incentivized review program we study failed to impact overall demand, despite the fact that the treatment affected the number and quality of reviews left for focal transactions. First, in many cases, even listings assigned to the control eventually obtained a first review. These reviews arrived quickly after experimental assignment and attenuated the differences in the distribution of first ratings received by listings in our experiment. Furthermore, since numerical ratings were averaged and were not visible until a listing received three reviews, it was difficult for one rating to appreciably change guests' perceptions of a listing. Finally, because Airbnb listing nights are not mass-produced goods, demand effects were only possible if the treatment could cause bookings for marginal nights.

5.3 The Effect of Incentivized Reviews on Transaction Quality

Although the incentivized review program failed to yield a statistically significant impact on overall demand, it's possible the program affected the quality of matches that occurred on Airbnb. This

[16] Rounding to half a star is a common design online and is used by Amazon, Etsy, and Yelp.
[17] Analysis of pricing is found in subsection B.5.

could be the case if, for instance, the presence of a review or the review text changed the decision of guests about which listing to pick. To test for this possibility, we construct transaction-level customer satisfaction proxies and compare the average level of each of these proxies for post-treatment transactions to control and treatment listings. More concretely, for each listing, l, we consider all transactions that occurred within 360 days of the focal stay checkout, did not have a cancellation, and had an observed payment. For this sample of transactions, we measure three customer satisfaction proxies: 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.[18]

Table 5 displays our results. Depending on the customer satisfaction proxy used, we find that the treatment had a neutral to negative effect on subsequent match quality. More specifically, we do not find a statistically significant impact of the treatment on the customer complaint rate (column 1), however this may stem from low statistical power due to the fact that customer complaints are extremely rare on the platform (approximately 1% of transactions result in a complaint). While the treatment had a positive impact on the review rate (column 2), this was caused at least in part by the fact that until treated listings received their first review, guests staying at those listings continued to receive the review incentive offer if they had not left a review within 8-9 days of checkout. Column 3 shows that conditional on a subsequent trip being reviewed, the rating was worse in the treatment group. This is consistent not only with the idea that subsequent match quality was worse in the treatment group, but also with the fact that in the treatment group subsequent trips may have resulted in incentivized reviews if the listing did not receive a review for the focal transaction (recall that incentivized reviews have lower star ratings on average).[19,20]

Although these analyses suggest that the incentivized review treatment caused subsequent transaction quality to decrease, they rely on customer customer satisfaction proxies such as guest

[18] 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).

[19] Hui et al. (2021) propose another reason that subsequent trips to treatment listings may have resulted in lower ratings on average: low ratings may be autocorrelated due to belief updating dynamics that affect review rates.

[20] Table C.3 shows that these results are not substantively affected by the inclusion of additional covariates in our regression model.

Table 5: 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.

ratings, which may be subject to a number of reporting biases. Using subsequent behavior as a proxy for customer satisfaction provides a stronger result that is more straightforward to interpret. Column 4 of Table 5 shows that guests to treated listings stayed for 1.6% fewer post-transaction nights compared to guests to control listings. This effect remains statistically significant, albeit smaller in magnitude, after including controls for guest and trip characteristics (column 5). The robustness of this result to the inclusion of these controls suggests that the apparent reduction in guests' subsequent platform usage is not because incentivized reviews induced matches with guests who had a lower baseline propensity to return to Airbnb. Instead, the simplest explanation for the results in columns 4 and 5 is that incentivized reviews induced *worse* matches and caused guests to use Airbnb less as a result. This could be the case if, for instance, incentivized five-star reviews were induced for low quality listings, causing these listings to appear to subsequent guests as higher quality than they actually were.[21]

We investigate whether this was in fact the case in our setting by considering how the predictive power of focal transaction ratings differs between the treatment and control groups.[22] If incentivized reviews were more inflated than non-incentivized reviews, then a five-star incentivized review should be less predictive of subsequent five-star non-incentivized reviews. If, on the other hand, incentivized reviews were *less* inflated then organic reviews, then incentivized reviews should be more predictive of subsequent five-star ratings. Table 6 displays the results of a regression where the outcome is the rating of a subsequent review and the explanatory variable is the star rating of the focal review interacted with the treatment. We see that treatment five-star ratings were less associated with future five star ratings than control five-star ratings, regardless of the sample used (the next review (column 1) or all subsequent reviews (column 2)) and whether covariates are added (columns 3 and 4). This result suggests that five-star ratings in the treatment were indeed more inflated than five-star ratings in the control.

[21] The specific dynamics that could cause an incentivized review program to lead to worse matches are formalized in Appendix A.

[22] Although this is the most natural way to measure inflation in review ratings, it requires conditioning on the first review that a listing received. Thus far, we have avoided conditioning on the incidence or star-rating of the focal review. Although conditioning on reviews has intuitive appeal, this approach to analysis suffers from severe omitted variable bias. Keeping this in mind, the analysis that follows should be interpreted with caution.

Table 6: Predictive Power of Ratings: Treatment vs Control

	Rating			
	(1)	(2)	(3)	(4)
Constant	4.093***	4.285***		
	(0.0509)	(0.0272)		
2 Star	0.1036	0.0347	0.0984	0.0276
	(0.0640)	(0.0367)	(0.0635)	(0.0351)
3 Star	0.2025***	0.0932**	0.2097***	0.0991***
	(0.0540)	(0.0297)	(0.0538)	(0.0285)
4 Star	0.3924***	0.2460***	0.4000***	0.2484***
	(0.0513)	(0.0276)	(0.0511)	(0.0266)
5 Star	0.5625***	0.3923***	0.5549***	0.3795***
	(0.0510)	(0.0273)	(0.0508)	(0.0263)
1 Star × Treatment	0.0691	-0.0101	0.0729	-0.0063
	(0.0636)	(0.0383)	(0.0631)	(0.0370)
2 Star × Treatment	-0.0680	-0.0752*	-0.0616	-0.0649*
	(0.0496)	(0.0330)	(0.0490)	(0.0315)
3 Star × Treatment	-0.0316	-0.0271·	-0.0252	-0.0222
	(0.0222)	(0.0144)	(0.0220)	(0.0138)
4 Star × Treatment	-0.0295***	-0.0322***	-0.0295***	-0.0298***
	(0.0082)	(0.0054)	(0.0081)	(0.0051)
5 Star × Treatment	-0.0199***	-0.0183***	-0.0196***	-0.0172***
	(0.0046)	(0.0029)	(0.0046)	(0.0027)
Sub-sample	Next Review	All Subsequent Reviews	Next Review	All Subsequent Reviews
Covariates	No	No	Yes	Yes
Observations	134,245	1,308,783	134,089	1,304,491
R^2	0.02917	0.02041	0.05705	0.05192

Notes: This table displays a regression where the star rating for stays after the focal stay is regressed on ratings for the focal stays and the treatment. Columns (1) and (3) use the sample of only the review following the review from the focal stay. Columns (2) and (4) use all subsequent reviews in the sample. Columns (3) and (4) include covariates for characteristics of subsequent stays. These are 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, guest age on the platform, guest region, checkout week, nights booked, and number of guests. Standard errors are clustered at the listing level in specifications (2) and (4).

In summary, although we find that the incentivized review program affected the number and type of reviews left by Airbnb guests, we do not find evidence that the program affected quantities sold or revenues for listings or the platform. We identify a number of reasons that this is the case, including market conditions on Airbnb, the design of Airbnb's reputation system, and capacity constraints for listings on the platform. Although we do not find effects of the treatment on overall demand, we *do* find that the treatment affected the quality of matches occurring on the platform. Specifically, match quality, as measured both by the star ratings that subsequent guests left and the rate at which guests returned to the platform, decreased. This is at least in part due to the fact that ratings left under the incentivized review treatment were more inflated than those left under the status quo.

6 Discussion

In this paper, we analyze an experiment on Airbnb to determine the extent to which an incentivized review policy can improve market outcomes such as demand and match quality. Although we find the incentivized review program was effective at inducing reviews, and that these reviews were more negative on average, the treatment did not affect overall demand and actually decreased the quality of subsequent matches on the platform.

The fact that we did not find an impact of the treatment on overall demand stands in contrast to Park, Shin and Xie (2021) and Vana and Lambrecht (2021), who both report large effects of reviews on demand. We argue that the structure of the market, the design of Airbnb's reputation system, and the capacity constraints of Airbnb listings are critical for understanding these findings. If sellers are expected to quickly accumulate reviews, then the effect of a marginal review is likely to be small. The effect of incentivized reviews is further reduced by the rounding of ratings to a coarse average. Both the speed at which sellers can accumulate reviews on Airbnb and the rounding of ratings in the platform's reputation system caused the incremental effects of one review on the platform to be small. Sellers on Airbnb also have limited capacity, putting a ceiling on the impact

that any intervention can have on quantities sold. One might worry that although Airbnb exhibits these qualities, other marketplaces may not. However, we also find evidence of these marketplace dynamics in a scraped dataset that describes reviewing behavior on a large home improvement services platform,[23] and suspect that many other marketplaces have similar dynamics.

A key goal of a reputation system is to create good matches between buyers and sellers. In contrast, we find that incentivized reviews caused worse matches on the platform — whether measured by subsequent ratings or by guest return rates to the platform. We argue that these worse matches were caused at least in part by the fact that incentivized ratings were more inflated conditional on a rating level. In particular, conditional on a given star rating, incentivized ratings represented a worse experience than non-incentivized ratings. This finding highlights that it is important to look beyond the ratings distribution when determining whether a policy increases or decreases review inflation.

Our negative evaluation of a specific incentivized review program does not preclude other interventions targeted at increasing review rates from having positive effects. Our treatment induced reviews for a specific set of transactions and had imperfect compliance — only 37.05% of treated transactions were reviewed. A policy that induced reviews for a different subset of transactions, for example those with customer complaints, could have different effects on market outcomes. Alternatively, a more *intensive* incentivized review policy that encourages reviews for many transactions per seller could have larger effects.

Lastly, the incentivized review policy that we study is not well-suited toward solving the cold-start problem in online marketplaces. In order to solve the cold-start problem, a platform would need to consider alternative interventions such as hiring new sellers directly as in Pallais (2014). Other policies that may solve the cold-start problem include subsidizing transactions with new sellers, 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.

[23] Appendix B.8 provides more detail on how this data was collected and on the analysis we conducted using it.

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.

Benson, Alan, Aaron Sojourner, and Akhmed Umyarov. 2020. "Can reputation discipline the gig economy? Experimental evidence from an online labor market." *Management Science*, 66(5): 1802–1825.

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

Brandes, Leif, David Godes, and Dina Mayzlin. 2019. "What drives extremity bias in online reviews? Theory and experimental evidence." *Theory and Experimental Evidence (September 10, 2019)*.

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.

Filippas, Apostolos, John Joseph Horton, and Joseph Golden. 2022. "Reputation inflation." *Marketing science*, 483–484.

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.

Garg, Nikhil, and Ramesh Johari. 2021. "Designing informative rating systems: Evidence from an online labor market." *Manufacturing & Service Operations Management*, 23(3): 589–605.

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

He, Sherry, Brett Hollenbeck, and Davide Proserpio. 2020. "The market for fake reviews." *Available at SSRN*.

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.

Hui, Xiang, Zekun Liu, and Weiqing Zhang. 2020. "Mitigating the Cold-start Problem in Reputation Systems: Evidence from a Field Experiment." *Available at SSRN*.

Ifrach, Bar, Costis Maglaras, Marco Scarsini, and Anna Zseleva. 2019. "Bayesian social learning from consumer reviews." *Operations Research*, 67(5): 1209–1221.

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.

Luca, Michael, and Georgios Zervas. 2016. "Fake it till you make it: Reputation, competition, and Yelp review fraud." *Management Science*, 62(12).

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

Mayzlin, Dina, Yaniv Dover, and Judith Chevalier. 2014. "Promotional reviews: An empirical investigation of online review manipulation." *American Economic Review*, 104(8): 2421–55.

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.

Papanastasiou, Yiangos, Kostas Bimpikis, and Nicos Savva. 2018. "Crowdsourcing exploration." *Management Science*, 64(4): 1727–1746.

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.

Vana, Prasad, and Anja Lambrecht. 2021. "The effect of individual online reviews on purchase likelihood." *Marketing Science*.

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.

Wu, Chunhua, Hai Che, Tat Y Chan, and Xianghua Lu. 2015. "The economic value of online reviews." *Marketing Science*, 34(5): 739–754.

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, where informativeness is a measure of 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 buyer's beliefs about ratings 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 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:

$$u_{il} = \theta_i + Q_l - p \tag{3}$$

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:

$$b_{il} = \arg\max_{b\in\{0,1\}} \mathbf{1}\{b=1\}(E_Q[\theta_i + Q_l - p|\Omega_l]) = \arg\max_{b\in\{0,1\}} \mathbf{1}\{b=1\}(\theta_i + q_i(\Omega_l) - p) \quad (4)$$

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.[24] 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:[25]

$$\begin{aligned} E_{\theta,Q,\Omega} =& \mu(1 - p + E_\theta[\theta]) \\ &+ (1-\mu)E_\theta[-(p-\theta)P_\Omega[q \geq p - \theta|Q=0]] \\ &+ \mu E_\theta[-(1-p+\theta)P_\Omega[q \leq p-\theta|Q=1]] \end{aligned} \quad (5)$$

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

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

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

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:

$$\begin{aligned}&(\tau_{H,1} + \tau_{L,1})Pr(p - q(0) > \theta > p - q(1)) - \\ &(\tau_{H,-1} + \tau_{L,-1})Pr(p - q(-1) > \theta > p - q(0))\end{aligned} \tag{6}$$

Equation 6 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 6.28 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 6 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 in the treatment system.

Ex-ante, the direction of the change in demand due to the shift from Ω_c to Ω_t in our setting is ambiguous. Our analysis in section 4 shows that the there is a much larger 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/or the decrease in demand due to negative reviews is large, in which case the overall effect of incentivized reviews on demand may be small (or negative). Furthermore,

the strength of both positive and negative demand effects depend on the extent to which an individual review updates buyer beliefs. Bayesian updating suggests that buyer beliefs about quality should be most affected in cases where no other reviews are present. By definition, our treatment targets Airbnb listings that do not have any prior reviews, but we would also expect the effects of incentivized reviews to be mediated by whether listings are able to quickly obtain other reviews. In subsection 5.2, we document that the median difference in time between a listing's first review in the control and treatment group is only 6 days.

The effects of incentivized reviews on expected utility 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, which 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 worse matches caused by incentivized reviews. The change in expected utility from incentivized reviews can be expressed as follows:

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

Equation 7 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.[26] 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.

Whether incentivized reviews increase or decrease expected utility (i.e., the net sum of all of these terms) depends on the composition (high or low quality) of non-reviewed listings for whom the incentive induces a review, and whether or not the induced reviews match the quality of the reviewd listings. This empirical question is explored in the main text of this paper.

[26] 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

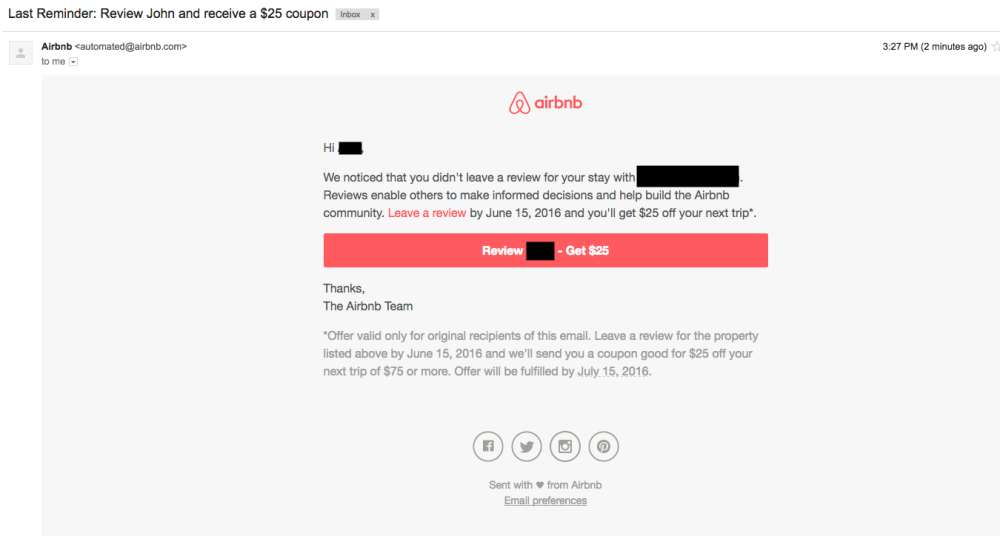
The number of days between the checkout and emails in the experiment was intended to be nine days for most of the sample. After March 29, 2016, the number of days within which a review must have been submitted to determine eligibility was changed to seven.

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.8 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.9 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

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.

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 estimate 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 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:

$$text\_pos_l = \beta_0 + \beta_1 T_l + \gamma_r + \epsilon_l \tag{8}$$

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 8. 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. 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 the time horizon at which we measure outcomes matters, 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 Figure C.10 we measure the intent to treat and local average treatment effects at differing time horizons. We find that the treatment effect on views and transactions gradually rises, stabilizes at 60 days after focal stay checkout, and starts falling after 120 days. In contrast, we find that the effects on nights and booking value remain close to 0 and not statistically significant across all

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.

time horizons.

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.2, 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.5 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 C.11 shows the effects of the treatment on overall views, and on views originating from a search. The effect on views from search was 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 increased in the treatment but the search ranking did not change. We conclude that the presence of information about reviews in search results mattered. Searchers saw that treated listings had more reviews and this induced them to click on the listing

page to view more information.

B.5 Do Incentivized Reviews Affect Prices?

In this section, we consider how the prices that hosts set were affected by incentivized reviews and whether incentivized reviews predict subsequent pricing decisions. We find that the treatment did not affect prices. We also find that ratings are strongly predictive of prices, but that both control and incentivized ratings have a similar correlation to prices.

To conduct this investigation, we use the sample of all transactions that occur within 360 days of checkout date of the focal transaction. We then regress the per-night price on the treatment status, the nightly price for the focal transaction, star ratings, and their interactions with the treatment. Table C.6 displays the results of these regressions. Column (1) shows that the treatment did not affect subsequent prices and that focal prices are strongly correlated with subsequent prices. Column (2) shows that listings that receive five star ratings had much higher subsequent prices than listings which do not. Lastly, column (3) shows that there is no statistically significant interaction effect between treatment status and star rating. In supplementary analysis, we also consider listed prices rather than transaction prices and find similar results (Table C.7).

B.6 Large Heterogeneous Treatment Effects Do Not Explain the Small Average Treatment Effects of Incentivized Reviews

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.

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)):[27]

$$y_l = \beta_0 + \beta_1 T_l + \beta_2 X_l + \beta_3 T_l (X_l - \bar{X}) + \epsilon_l \tag{9}$$

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 B.2 displays the results from Equation 9. 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

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

Table B.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)
Observations	640,936	640,936	640,936
R^2	0.16055	0.13454	0.18840

Notes: This table displays the regression estimates from Equation 9, 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 B.6. 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.

the platform does not explain the small average effects of the treatment.[28]

B.7 Treatment Effect Heterogeneity by Review Rating

Next, we investigate whether heterogeneous effects 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.12 displays the observational estimates of the effect of a review in the control group on 120 day nights and revenue. We see that listings with a first transaction that receives a five star review have much more demand than listings where the first transaction is not reviewed, 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 shrink-

[28] We also conduct a more standard analysis of heterogeneity in Table C.4.

age 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 effect.[29]

We have failed to find that heterogeneity in the effects of incentivized reviews 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 were typically small and that naive observational estimates of the effects of reviews were mostly explained by selection bias.

B.8 Home Improvement Platform Scrape

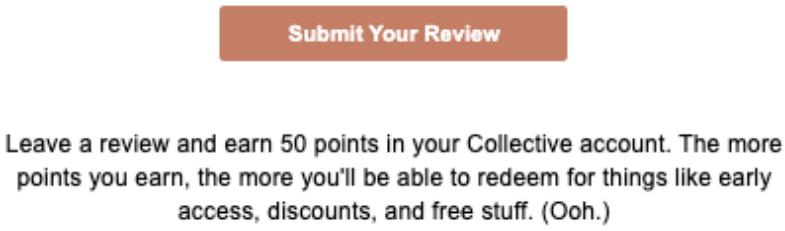
In 2018, Farronato et al. (2020) performed a comprehensive web-crawl of a large home improvement services platform. They identified the largest three cities for each state in terms of unique home improvement professionals in categories subject to licensing, and joined that list with the top 100 cities 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 6-day difference in the median time to first review between our control and treatment groups on 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.

[29] 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.

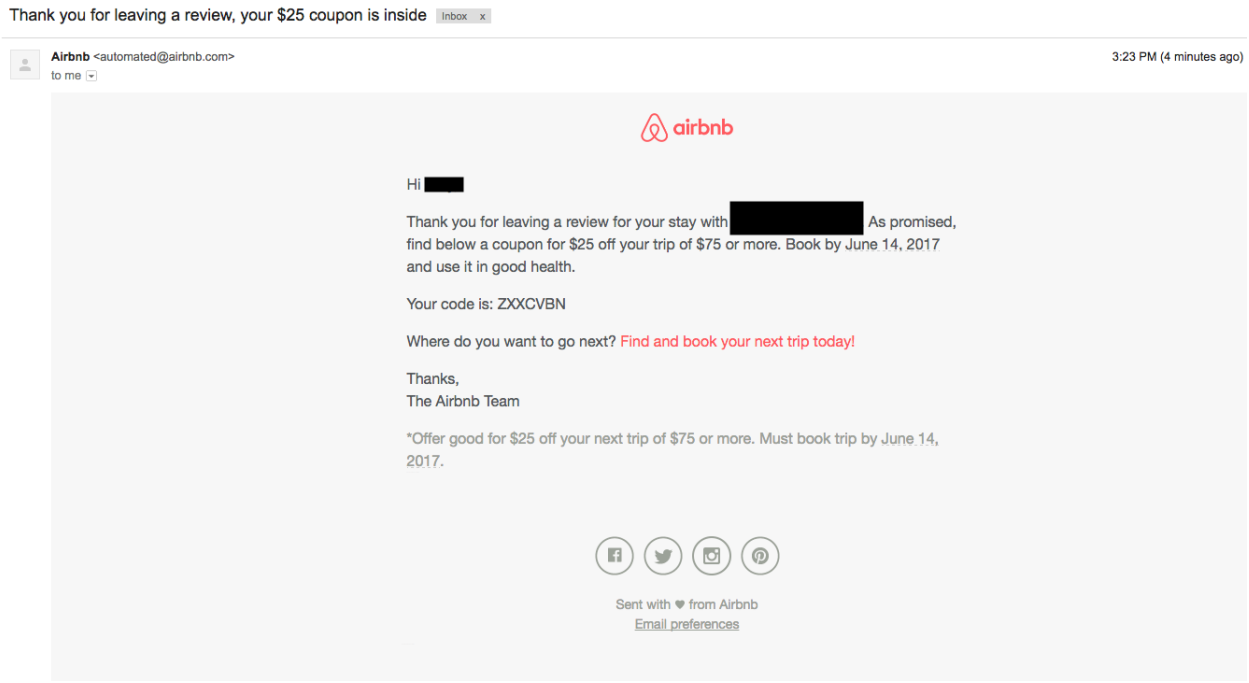
C Appendix: Additional Figures and Tables

Figure C.1: Incentivized Review Solicitation from the Girlfriend Collective



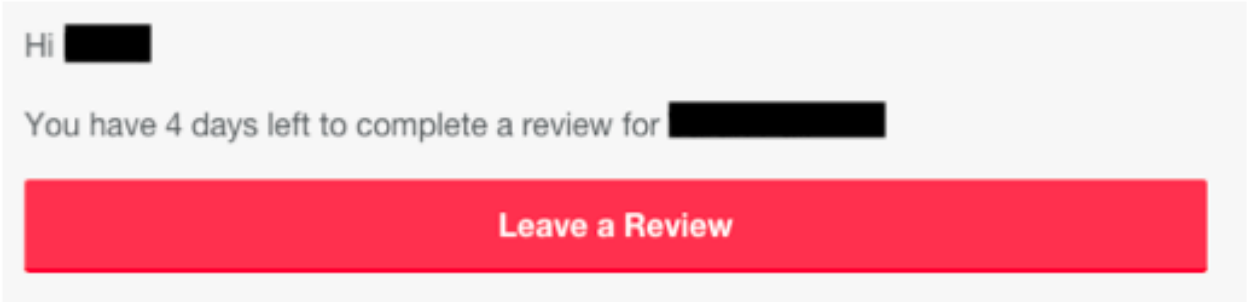
Notes: This figure displays an incentivized review email sent by the Girlfriend Collective.

Figure C.2: Email Sent After Incentivized Review



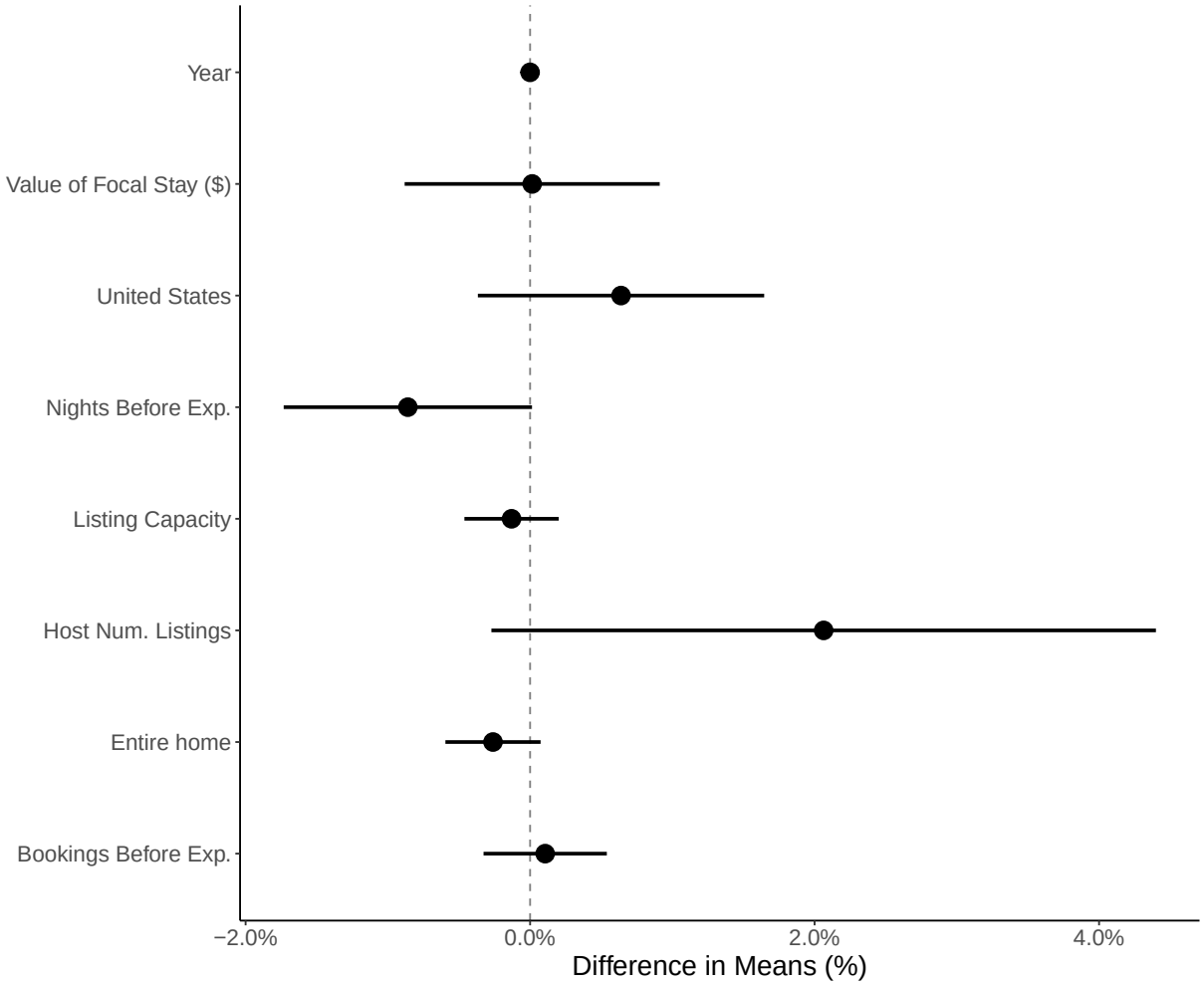
Notes: Displays the email sent to guests who had stayed in treatment listings that had not yet received a review on Airbnb after a certain number of days, issuing them a coupon after leaving a review.

Figure C.3: Email Sent to the Control Group



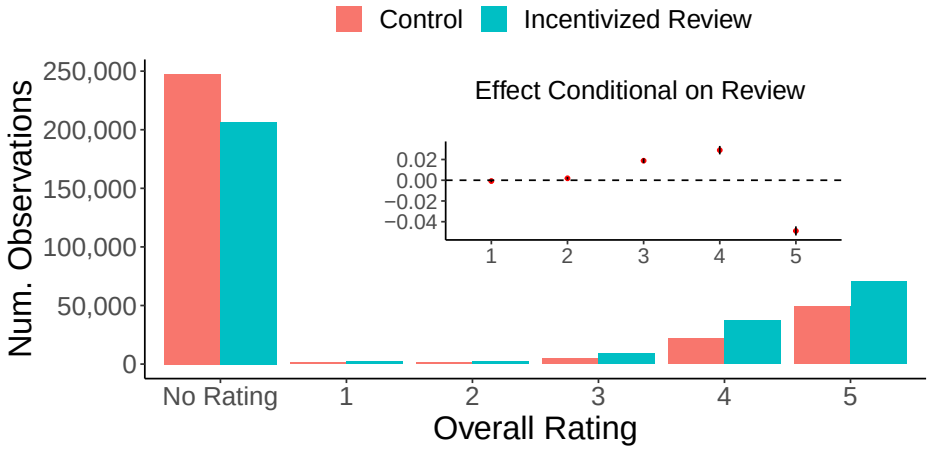
Notes: Displays the email sent to guests who had stayed in control listings that had not yet received a review on Airbnb after a certain number of days.

Figure C.4: Balance Assessment for Experiment



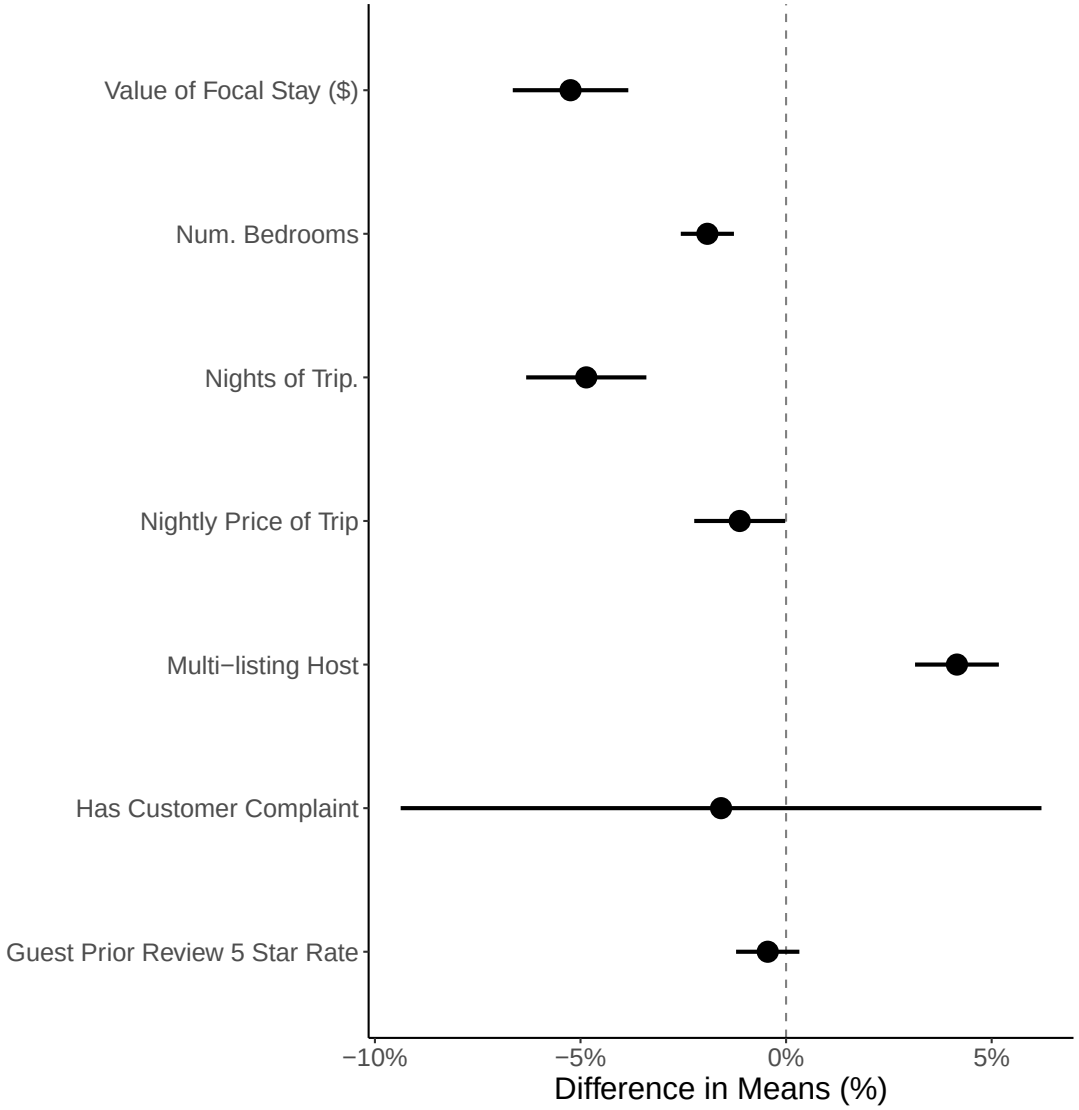
Notes: This plot displays the difference in means between the treatment and control groups for pre-treatment covariates and the days between checkout and email date. No differences were statistically significant in the pre-treatment covariates (year of stay, dollar value of transaction, whether the listing 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.5: Distribution of Ratings for Focal Stay



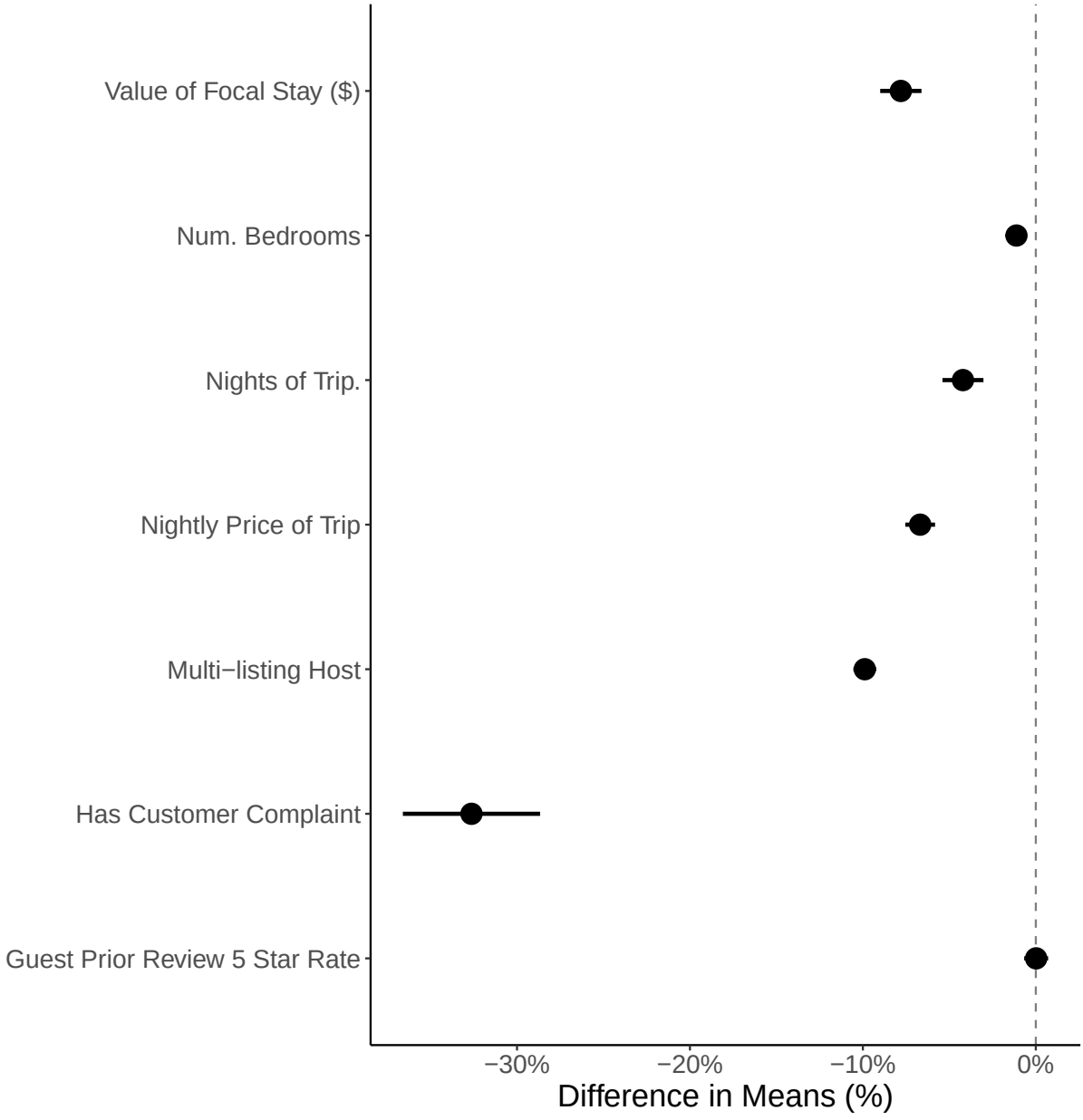
Notes: Comparison of the distribution of ratings left in the treatment group and the control group during the experiment. We only include the first review left for each listing. The inset plot contains the treatment effect and 95% confidence interval conditional on a rating being submitted.

Figure C.6: Differences in Characteristics of Reviewed Transactions
Treatment vs Control



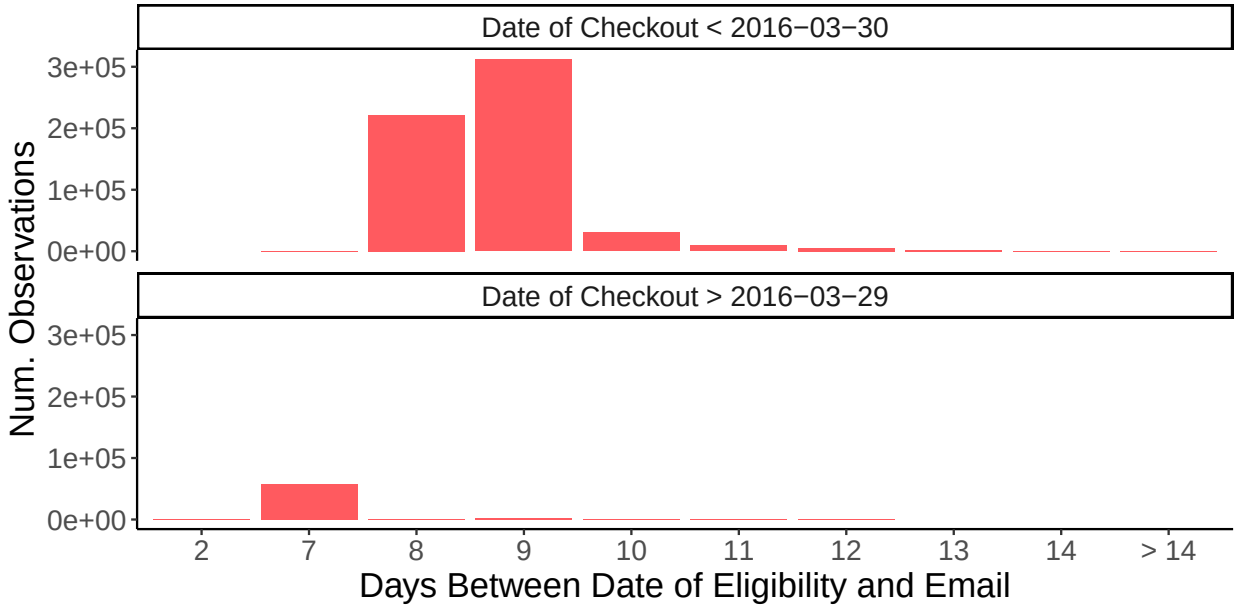
Notes: This plot displays the difference in means between the treatment and control groups for trip characteristics.

Figure C.7: Differences in Characteristics of Transactions
Reviewed vs Non-Reviewed



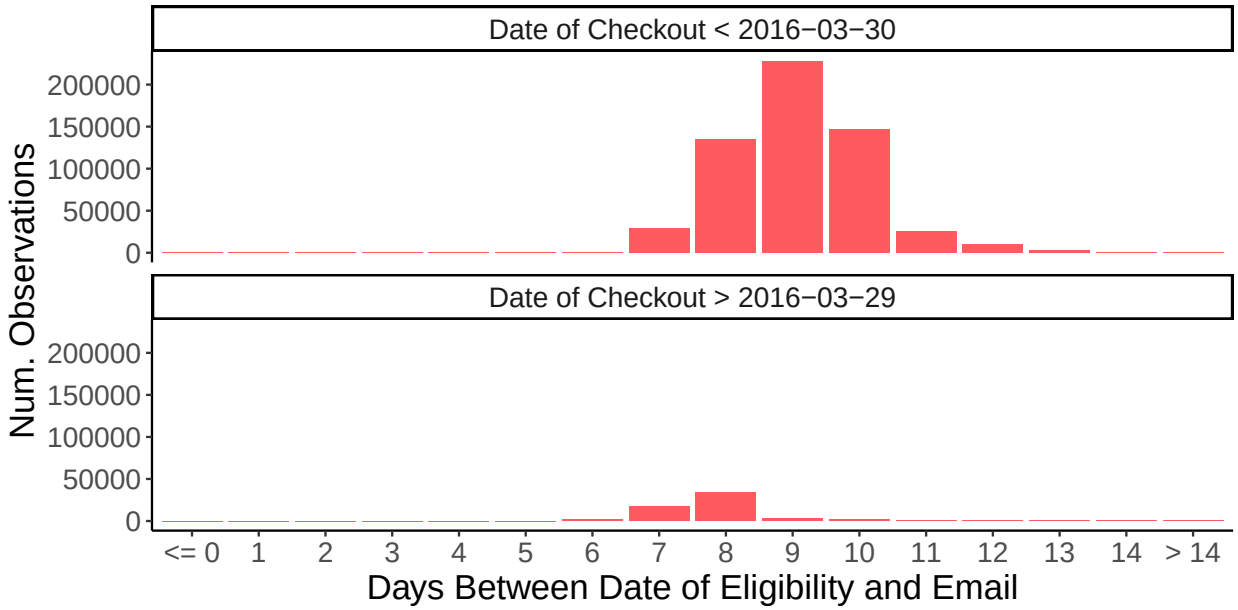
Notes: This plot displays the difference in means between reviewed and non-reviewed transactions in the treatment group.

Figure C.8: Days Between Assigned Date and Email



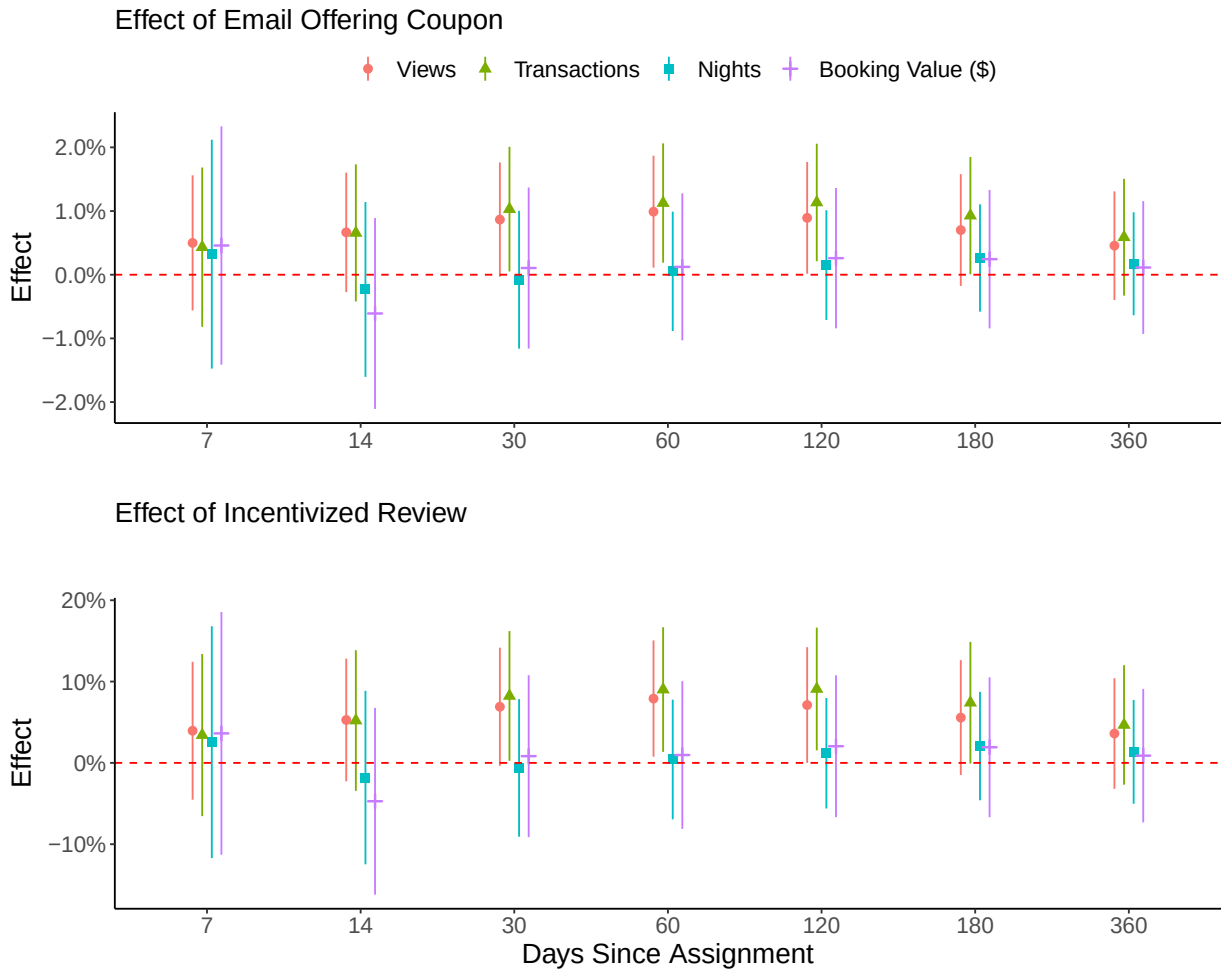
Notes: This figure plots the histogram of days between email and the assigned checkout used by the email dispatch system. Note that no email was logged for 2.2% of observations, either due to missing logging or email dispatch errors.

Figure C.9: Days Between Realized Checkout and Email



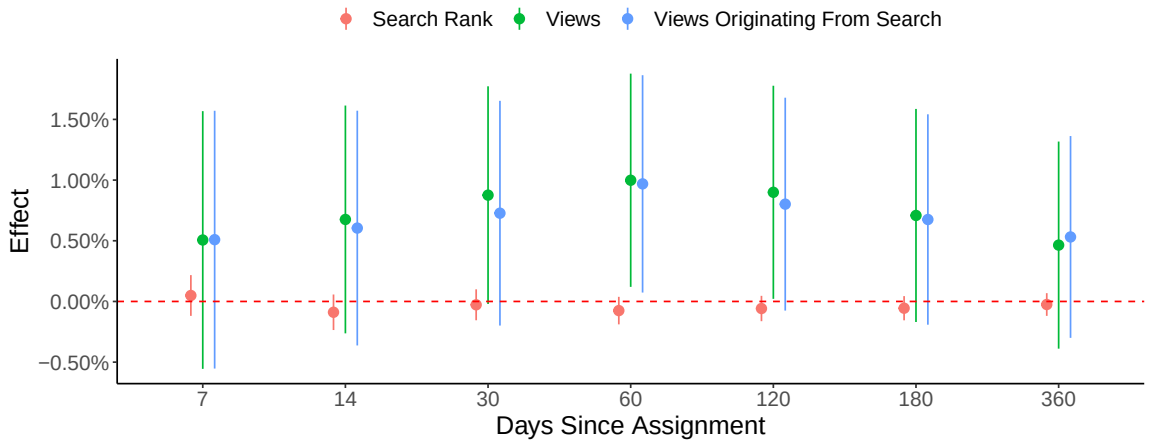
Notes: This figure plots the histogram of days between email and the realized checkout of the focal transaction. Note that no email was logged for 2.2% of observations, either due to missing logging or email dispatch errors.

Figure C.10: Cumulative Effects of Treatment on Listing Outcomes



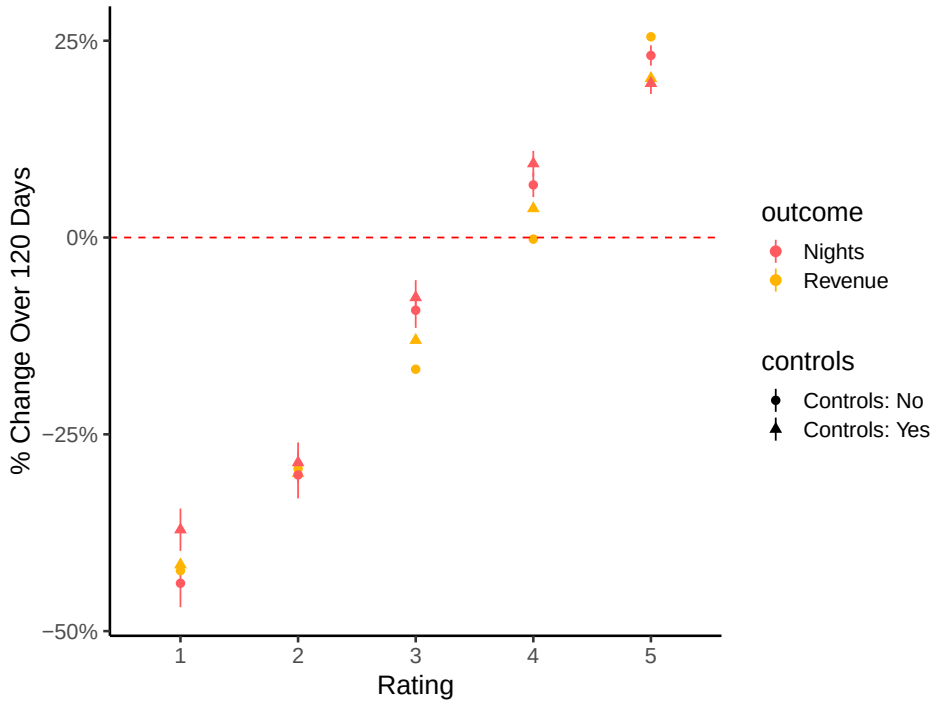
Notes: The figure plots the effects and 95% confidence intervals from Equation 1, where coefficients are transformed into percent terms by dividing by the intercept. Each point represents the effect of a listing's guest receiving a treatment email on an outcome measured cumulatively between the checkout 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 C.11: 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.

Figure C.12: 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: Intensive Margin Regression

	Has Res.	Num. Res.	Has Res.	Num. Res.	Has Res.	Num. Res.
	2 Months After		**4 Months After**		**12 Months After**	
	(1)	(2)	(3)	(4)	(5)	(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)
Sub-sample	All	Cond. on Res.	All	Cond. on Res.	All	Cond. on Res.
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.3: Effects of Treatment on Transaction Quality - With Covariates

	Complaint	Reviewed	Star Rating
	(1)	(2)	(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.4: 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.5: 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.

Table C.6: Ratings and Transaction Prices

	Log(Subsequent Price)		
	(1)	(2)	(3)
Constant	0.6589***	0.6530***	0.6522***
	(0.0068)	(0.0068)	(0.0067)
Treatment	3.26×10^{-5}	-0.0019	-0.0002
	(0.0015)	(0.0016)	(0.0019)
Log(Focal Price)	0.8581***	0.8581***	0.8581***
	(0.0015)	(0.0015)	(0.0015)
1 Star		0.0098	0.0186
		(0.0152)	(0.0252)
2 Star		-0.0364**	-0.0340*
		(0.0113)	(0.0158)
3 Star		-0.0213***	-0.0084
		(0.0056)	(0.0093)
4 Star		-0.0056*	-0.0066
		(0.0025)	(0.0039)
5 Star		0.0386***	0.0422***
		(0.0021)	(0.0033)
Treatment × 1 Star			-0.0149
			(0.0316)
Treatment × 2 Star			-0.0041
			(0.0219)
Treatment × 3 Star			-0.0197
			(0.0117)
Treatment × 4 Star			0.0010
			(0.0051)
Treatment × 5 Star			-0.0066
			(0.0043)
R^2	0.76434	0.76483	0.76484
Observations	2,389,144	2,389,144	2,389,144

Notes: This table displays regressions measuring the correlation between focal transaction rating and subsequent nightly transaction prices set by sellers. 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. 'Treatment' is an indicator for whether the lsiting was assigned to the treatment, 'Focal price' is the nightly price of the focal transaction (inclusive of transaction and cleaning fees), and 'Star' corresponds to a rating for the focal transaction. Note that the omitted category represents cases when there was no review for the focal transaction. Standard errors are clustered at a listing level.

Table C.7: Ratings and 2017 Listed Prices

	Log(2017 Listed Price)		
	(1)	(2)	(3)
Constant	0.6989***	0.6957***	0.6948***
	(0.0050)	(0.0050)	(0.0050)
Log(Focal Nightly Price)	0.8216***	0.8213***	0.8212***
	(0.0011)	(0.0011)	(0.0011)
Treatment	0.0005	-0.0004	0.0017
	(0.0014)	(0.0014)	(0.0017)
1 Star		0.0075	0.0286
		(0.0132)	(0.0221)
2 Star		-0.0486***	-0.0427*
		(0.0110)	(0.0190)
3 Star		-0.0465***	-0.0371***
		(0.0051)	(0.0086)
4 Star		-0.0194***	-0.0166***
		(0.0023)	(0.0037)
5 Star		0.0401***	0.0435***
		(0.0018)	(0.0027)
1 Star × Treatment			-0.0361
			(0.0274)
2 Star × Treatment			-0.0101
			(0.0232)
3 Star × Treatment			-0.0149
			(0.0107)
4 Star × Treatment			-0.0050
			(0.0048)
5 Star × Treatment			-0.0063
			(0.0036)
R^2	0.74541	0.74600	0.74601
Observations	338,376	338,376	338,376

Notes: This table displays regressions measuring the correlation between focal transaction rating and log of the posted prices on January 1, of 2017. 'Treatment' is an indicator for whether the listing was assigned to the treatment, 'Focal price' is the nightly price of the focal transaction (inclusive of transaction and cleaning fees), and 'Star' corresponds to a rating for the focal transaction. Note that the omitted category represents cases when there was no review for the focal transaction.

Table C.8: Active Status in 2017

	Active in 2017		
	(1)	(2)	(3)
Constant	0.5163***	0.5046***	0.5018***
	(0.0009)	(0.0009)	(0.0010)
Treatment	0.0015	-0.0036**	0.0025
	(0.0012)	(0.0012)	(0.0015)
1 Star		-0.1800***	-0.1770***
		(0.0080)	(0.0126)
2 Star		-0.1123***	-0.1050***
		(0.0078)	(0.0128)
3 Star		-0.0432***	-0.0271***
		(0.0043)	(0.0075)
4 Star		0.0413***	0.0525***
		(0.0022)	(0.0035)
5 Star		0.0713***	0.0827***
		(0.0016)	(0.0024)
1 Star × Treatment			-0.0065
			(0.0164)
2 Star × Treatment			-0.0134
			(0.0161)
3 Star × Treatment			-0.0260**
			(0.0091)
4 Star × Treatment			-0.0196***
			(0.0045)
5 Star × Treatment			-0.0208***
			(0.0033)
R^2	2.23×10^{-6}	0.00465	0.00474
Observations	654,595	654,595	654,595

Notes: This table displays regressions measuring the correlation between focal transaction rating and whether a listing was active on January 1, 2017. 'Treatment' is an indicator for whether the listing was assigned to the treatment and 'Star' corresponds to a rating for the focal transaction. Note that the omitted category represents cases when there was no review for the focal transaction.