Appendix
Reciprocity and Unveiling in Two-sided Reputation Systems: Evidence from an Experiment on Airbnb

Andrey Fradkin*[1], Elena Grewal†[2], and David Holtz‡[3]

[1]Boston University and MIT Initiative on the Digital Economy
[2]Data 2 The People
[3]MIT Sloan School of Management

June 3, 2021

A Logging of Reviews

In this section we discuss several details about the logging of review and treatment data in our sample. Overall ratings and review text were required and logged more than 99.9% of the time conditional on a guest review. Whether the other ratings were required depends on the device that was used to submit the review. On iOS, the sub-ratings and recommendations were required. On a desktop browser, the sub-ratings and recommendations were not required and are missing for 5.5% of guest reviews and 4.4% of host reviews. On Android, the sub-ratings were required but the anonymous recommendation was not logged. 79% of guest reviews and 76% of host reviews were submitted via a desktop browser in our sample.

The simultaneous reveal review experiment launched on May 8, 2014 and our sample includes trips with checkout dates between May 7, 2014 and June 12, 2014. However, there were two logging issues during the experiment.

The first logging issue occurred at the outset of the experiment. When launched on May 8, Airbnb's experiment logging framework had bugs. These were fixed by May 11, 2014. Our main analysis sample simply excludes transactions with checkout dates earlier than May 10, 2014. However, if being exposed to the treatment between May 8 and May 11 affected subsequent trips, this could impact our analysis. To verify that this is not the case, we create a new sample that excludes any host with a trip ending on May 7, May 8, or May 9. Note that this sample excludes more active hosts, who are more likely to have a transaction ending on any given day. Figure A5 displays the

*Primary Author: fradkin@bu.edu
†Primary Experiment Designer
‡dholtz@mit.edu

baseline experimental results for this sample. The treatment effects in the two samples are similar in magnitude and precision.

A second logging issue occurred towards the end of our experiment. Treatment assignment logs are missing for some transactions on June 6 and June 7. We account for this issue with the following procedure. For hosts whose first transaction treatment assignment is missing because it ends on one of these days, we exclude the host from the sample. We keep transactions for hosts whose first transaction is after the June 7 because we can observe treatment assignment.

B Measuring Review Text

The text of a review is the most publicly salient type of the information collected by the review system because the text of a review is permanently associated with an individual reviewer. In this paper, we focus on the sentiment of the text, e.g. whether the text contains only positive information or whether it includes negative phrases and qualifications. We use a regularized logistic regression, a common technique in machine learning, to classify the review text based on the words and phrases that appear in the text.

In order to train a classifier, we need "ground truth" labeled examples of both positive and negative reviews. We select a sample of reviews that are highly likely to be either positive or negative based on the ratings that guests submitted. Reviews by guests that we use as positive examples for guests and hosts are ones that have five star ratings. Reviews by guests that are examples of negative reviews are ones with a one or two star rating. Reviews by hosts that are examples of negative reviews are ones which have either a non-recommendation or a sub-rating lower than four stars. Foreign language reviews were excluded from the sample.

We use reviews submitted between January 2013 and March 2014. Because positive reviews are much more common than negative reviews, the classification problem would be unbalanced if we used the entire sample. Therefore, we randomly select 100,000 examples for both positive and negative reviews. Once we obtain these samples, we remove special characters in the text such as punctuation and we remove common "stop words" such as "a" and "that".[1] Each review is transformed into a vector for which each entry represents the presence of a word or phrase (bigrams and trigrams), where only words that occur at least 300 times are included. We tested various thresholds and regularizations to determine this configuration.

We evaluate model accuracy in several ways. First, we look at the confusion matrix describing model predictions on a 20% hold out sample. For guest reviews of listings, 19% of reviews with low ratings were classified as positive and 9% of reviews with high ratings were classified as negative. The relatively high rate of false positives reflects not only predictive error but the fact that some guests misreport their true negative experiences. We also evaluate model accuracy by doing a 10-fold cross-validation. The mean out of sample accuracy for our preferred model is 87%. Figures A6 and A7 display the most common phrases associated with negative reviews by guests and hosts and the relative frequency with which they show up in positive versus negative reviews. Phrases that commonly show up in negative reviews by guests concern cleanliness ('was dirty'), smell ('musty'), unsuitable furniture ('curtains'), noise ('loud'), and sentiment ('acceptable') and

[1] These words are commonly removed in natural language applications because they are thought to contain minimal information.

phrases that commonly show up in negative reviews by hosts include ('would not recommend'), ('rude'), and ('smoke').

C Experimental Validity and Additional Results

Table AI displays the balance of observable characteristics in the experiments. There are no statistically significant differences in characteristics between the treatment and control guests or listings in the simultaneous reveal experiment.

Table AII displays the control and treatment means, as well as the treatment effects, for the following review related variables: whether the overall rating was 5 stars, whether all category ratings were 5 stars, and whether the review text was classified as negative.

D Linear Model of Review Timing Effects

In this section we discuss an alternative way to study the desire to reveal review information using a linear model. We would like to measure the effect of receiving a review on the submission of reviews. In this procedure, the outcome variable is whether a review by a user comes within a day after a review by the counterparty. The sample is the set of of observations for which the focal user (guest or host) does not review first. This includes observations where the focal user does not review at all.

Table AIII displays the results from this model. We find that the treatment increases the probability of reviews within a day by 39% for guests and 74% for hosts. These effects persist when adding time of first review fixed effects (columns (2) and (4)). Lastly, in column (5), we interact the treatment with host prior reviews. We do not find substantial or statistically significant heterogeneity in the effect by host experience.

E Principal Stratification Details

In this section, we briefly describe the principal stratification method used to separate the treatment effects we observe into treatment effects on two distinct subpopulations: always reviewers (i.e., users who would write a review whether in the control or treatment arm of our experiment) and compliers (i.e., users who review their counterparty when enrolled in the treatment, but would not review their counterparty when enrolled in the control). A more detailed description of the principal stratification approach can be found in Ding and Lu (2017).

We first compute the probability that each user in our sample is a complier, always reviewer, or never reviewer. We accomplish this by using the marginal method described by Feller, Mealli and Miratrix (2017).[2] Under the principal stratification approach's monotonicity assumption, we can assume that non-reviewers in the treatment group are never reviewers, and reviewers in the control group are always reviewers. For all other users in the sample, we can estimate the probability that they are an always reviewer using a logistic regression model that is trained on data from the control group and predicts the choice to review using a set of user- and trip-level covariates. Similarly, we can estimate the probability that each of these users is a never reviewer using a logistic regression model that is trained on data from the treatment group and predicts the choice to review using the

same set of user- and trip-level covariates. In both cases, we predict the choice to review using the following covariates:

- Whether the guest has any prior trips
- Whether the guest has submitted a review before
- Whether the host has any prior trips
- Whether the host has submitted a review before
- Whether the guest has submitted text before
- The average text sentiment of prior guest reviews
- The average overall star rating of prior guest reviews
- Whether the host has an effective positive percentage (EPP)
- The host's EPP
- Whether the host manages many listings
- Whether the guest has a gender
- Whether the guest has any prior customer service tickets
- Whether the host has any prior customer service tickets
- The property type of the listing
- Whether the guest is from the US
- The log of the listing price
- Whether the booking was made with instant book

Once we have estimated the probability that each user is an always reviewer and never reviewer, we can calculate the probability that each user is a complier, since $P(complier)_i = 1 - P(always\,r)_i - P(never\,r)_i$. In cases where $P(always\,r)_i + P(never\,r)_i > 1$, we set $P(complier) = 0$ and normalize the probabilities that the user is an always reviewer or never reviewer so that they sum to 1. After estimating the probability that each user belongs to each stratum, we use these probabilities as weights to construct causal stratum-level treatment effect estimators. Point estimates and confidence intervals are calculated using the bootstrap ($n = 1000$). We use the 'basic' bootstrap confidence interval method from the function 'boot.ci' in R.

[2]We also estimate the probability that each user belongs to each stratum using the EM algorithm described by Ding and Lu (2017). However, in order to make the calculation of bootstrap standard errors computationally tractable, we conduct the majority of our analysis using probabilities obtained through the marginal method. The point estimates we obtain using the EM algorithm are qualitatively similar to those obtained with the marginal method.

We test that the principal stratification model that we have proposed is accurate using the balancing conditions proposed by Ding and Lu (2017). Simply put, the balancing conditions require that within each stratum, the treatment should not appear to have a causal effect on any function of the pretreatment covariates used to estimate a given unit's stratum. We estimate the effect of the treatment on each pretreatment covariate in each stratum. The estimated effects are nearly zero (with a maximum absolute value of 8.07×10^{-7}) across all strata and covariates, indicating that the balancing conditions are satisfied.

F Additional concerns with the experimental results

F.1 Changes in email text over the course of the experiment

One concern with our experiment is that the exact email copy sent to users varied over the course of the experiment. We do not have internal data about which user got which email and even about the universe of emails sent.

To investigate whether these changes in email text were important, we solicited Airbnb review emails from this time period via social media. For guests in the control group, we found three versions of the email, which varied in the color scheme and logo.[3] We believe that the difference in logo is due to a dynamic link in the email and that the users saw the old logo when they actually received the email during the experiment period.

Similarly, we found that Airbnb inserted an additional piece of content in some of the initial treatment emails sent to hosts (the exact time at which this began is unclear to us). This content describes how reviews have changed (Figure A8) and was deployed for some members of the treatment group.

We know that Airbnb was not randomizing the specific email copy concurrently with our experiment. Therefore, any changes to the email copy must have occurred over time, with a change on a particular date. To test whether these changes in the email copy were material, we estimate the effects of the treatment across the days of our experiment. The results are displayed for guests in Figure A9 and hosts in Figure A10. In both cases, there are no trends in the treatment effect over time and the effect is similar in magnitude across the days. This confirms that any changes to the email copy during our experiment did not have large effects on the treatment effect.

As a final comment, to the extent that the email copy changed during our experiment, our treatment effects reflects a mix of the email copy that was sent to different users. This, if anything, improves the external validity of our estimates since there are many ways a platform could inform users about a new reviewing policy and other platforms may do so in a way which is more similar to one or the other email sent by Airbnb.[4]

F.2 Learning about the treatment

One concern with our interpretation of the experimental treatment effects is that users may not immediately learn about the change to the reputation system. For example, users may not have noticed that reviews had changed or that the change in reviews allowed them to be more honest

[3] Airbnb introduced a new logo and color scheme on July 16, 2014, which is after our experiment concluded.

[4] See Yarkoni (2019) for a discussion about the importance of using multiple stimuli for generalization.

when reviewing. This would attenuate the effects that we detect in our sample, but would not reverse our findings regarding the desire to reveal review information and reduced reciprocity. We provide evidence that learning effects were not of first order importance.

Figure A11 displays the review rates for guests and hosts over time, by treatment group. We see that following the end of the experiment, when all groups were assigned the treatment, the review rates in the control groups quickly jump to match the review rates in the treatment group and the review rates of the treatment group do not jump. Therefore, the longer exposure time for the treatment group did not have first-order consequences for reviewing rates. This shows that learning by users over the course of the experiment did not substantially affect review rates. It also suggests that the platform-wide launch of the policy did not result in effects larger than those predicted by the experimental treatment effects on review rates.

In A12 and A13 we plot how ratings evolved following the launch of simultaneous reveal to the entire site in July of 2014. We find that the differences in reviewing behavior following the launch are consistent with our observed treatment effects. Namely, the share of reviews with five star ratings falls and the share of reviews with 3 and 4 star ratings increases. Our results about both the review rates and ratings shows that the potential of learning over time about the policy does not overturn our main results.

References

Ding, Peng, and Jiannan Lu. 2017. "Principal Stratification Analysis Using Principal Scores." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3): 757–777.

Feller, Avi, Fabrizia Mealli, and Luke Miratrix. 2017. "Principal score methods: Assumptions, extensions, and practical considerations." *Journal of Educational and Behavioral Statistics*, 42(6): 726–758.

Yarkoni, Tal. 2019. "The Generalizability Crisis."

G Additional Tables

Table AI: Experimental Validity Check

Variable	Difference	Mean Treatment	Mean Control	P-Value	Stars
Total Bookings by Guest	-0.024	2.999	3.024	0.270	
US Guest	-0.002	0.285	0.286	0.558	
Guest Tenure (Days)	-2.065	268.966	271.032	0.271	
Host Listings	0.015	1.858	1.843	0.566	
Listing Reviews	-0.039	10.662	10.700	0.715	
Listing Trips Finished	-0.099	15.091	15.190	0.510	
US Host	0.002	0.266	0.264	0.547	
Multi-Listing	0.002	0.082	0.081	0.262	
Entire Property	-0.001	0.671	0.672	0.682	
Nights	-0.073	5.504	5.577	0.188	
Guests	-0.010	2.360	2.370	0.251	
Price Per Night	-3.138	291.690	294.828	0.273	
Observations	0.001			0.601	

This table displays the averages of variables in the treatment and control groups, as well as the statistical significance of the difference in averages between treatment and control. p<0.10, * p<0.05, ** p<0.01

Table AII: Other Experimental Treatment Effects

	Guest			*Host*		
	Control Mean	Treatment Mean	Effect	Control Mean	Treatment Mean	Effect
Overall Rating = 5	0.74	0.73	-0.01 ***			
All Ratings = 5	0.48	0.47	-0.01 ***	0.82	0.81	-0.01 ***
Review Text Negative	0.13	0.14	0.01 ***	0.03	0.03	1.8e-03 *

This table displays mean outcomes in the control and treatment, as well as treatment effects. The rating related outcomes are computed conditional on a review. The effect is displayed in percentage points. *p<0.1;**p<0.05;***p<.01;

Table AIII: Effects of Treatment on Reviewing Within a Day

	Dependent variable:				
	Reviews Within a Day of First Review				
	Guest		Host		
	(1)	(2)	(3)	(4)	(5)
Treatment	0.024***	0.023***	0.066***	0.064***	0.057***
	(0.002)	(0.002)	(0.003)	(0.003)	(0.006)
Treat * 1 - 3 Reviews					0.013*
					(0.008)
Treat * 4 - 12 Reviews					0.012
					(0.008)
Treat * $>$ 13 Reviews					0.001
					(0.008)
Mean of Y	0.062	0.062	0.089	0.089	0.089
Days Since Checkout FE	No	Yes	No	Yes	Yes
Observations	60,526	60,526	41,563	41,563	41,563
R^2	0.002	0.005	0.014	0.022	0.022
Note:					*p<0.1; **p<0.05; ***p<0.01

This table displays estimates from a linear probability model where the outcome is whether the guest (columns 1 - 2) or the host (columns 3 - 5) submitted a review within one day after the counterparty. Columns 2, 4, and 5 include fixed effects for the days since checkout of the initial review. 'Reviews' in column (5) refer to the number of reviews for the listing at the time of the booking.

Table AIV: Long-term Guest Outcomes

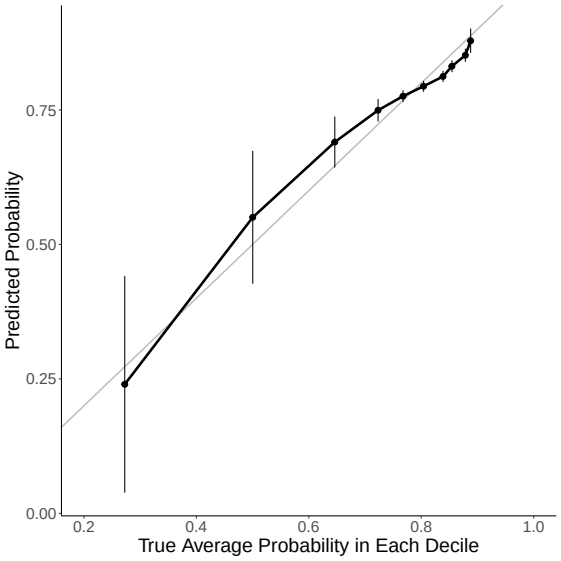
	Dependent variable:			
	Log(Nights in Exp.)	Log(Trips in Exp.)	Log(Nights by 2015)	Log(Bookings by 2015)
	(1)	(2)	(3)	(4)
Treatment	−0.006	−0.004*	−0.010*	−0.007
	(0.004)	(0.002)	(0.006)	(0.005)
Controls Included	Yes	Yes	Yes	Yes
Observations	115,157	115,157	115,157	115,157
R^2	0.065	0.078	0.186	0.047
Note:				*p<0.1; **p<0.05; ***p<0.01

This regression displays the effects of the treatment on the subsequent Airbnb usage of guests. The outcomes are the log of nights and trips taken during the experimental period as well as the log of nights and bookings which happened before 2015. Controls for guest market of origin, a time trend, the effective positive percentage of the listing, the log of the first price, and the number of reviews of the listing are included. Removing controls does not substantively affect the point estimates.

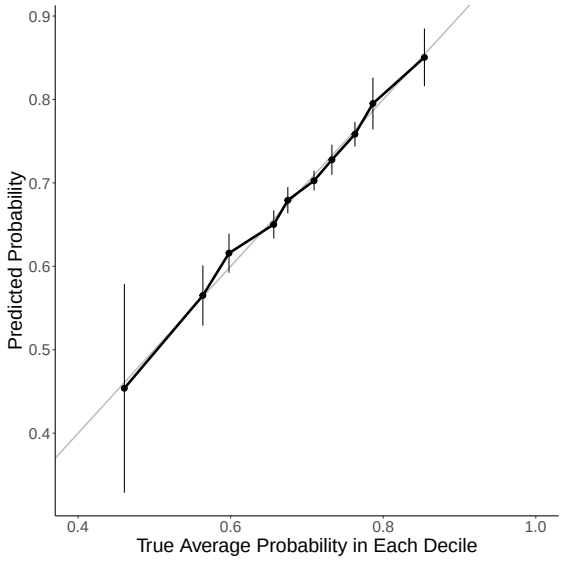
H Additional Figures

Figure A1: Calibration plot of review prediction in control group

(a) Host Reviews of Guest

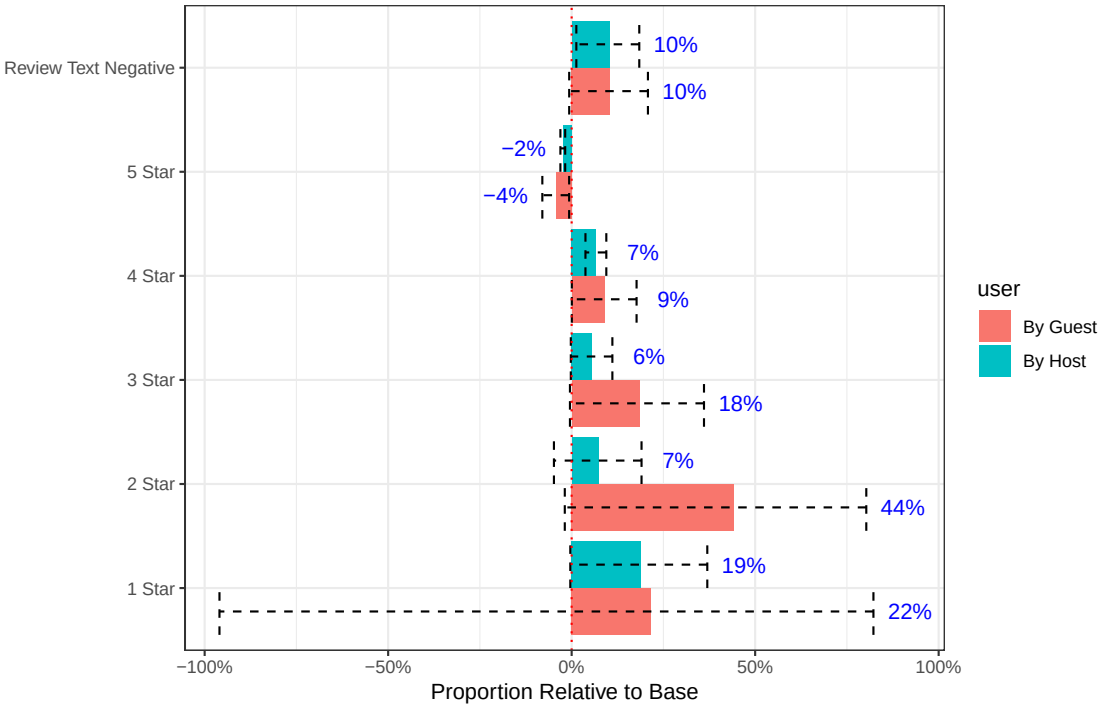


(b) Guest Reviews of Host



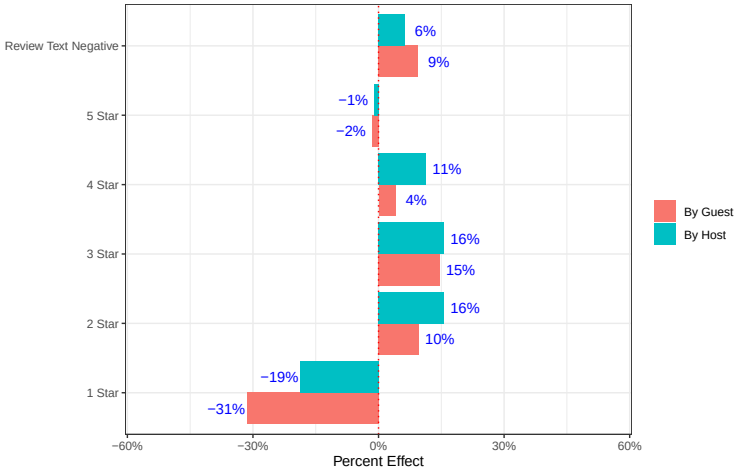
This figure plots the average review probability in the control group against the model predicted probabilities (y-axis). The line range represents the 95% range of predicted probabilities. The data is grouped into deciles of predicted probabilities so that each bin has approximately the same number of observations. The model used for prediction is described in Appendix E and 10-folds cross-validation is used to make the prediction out-of-sample.

Figure A2: Selection Into Reviewing - Complier Causal Effects



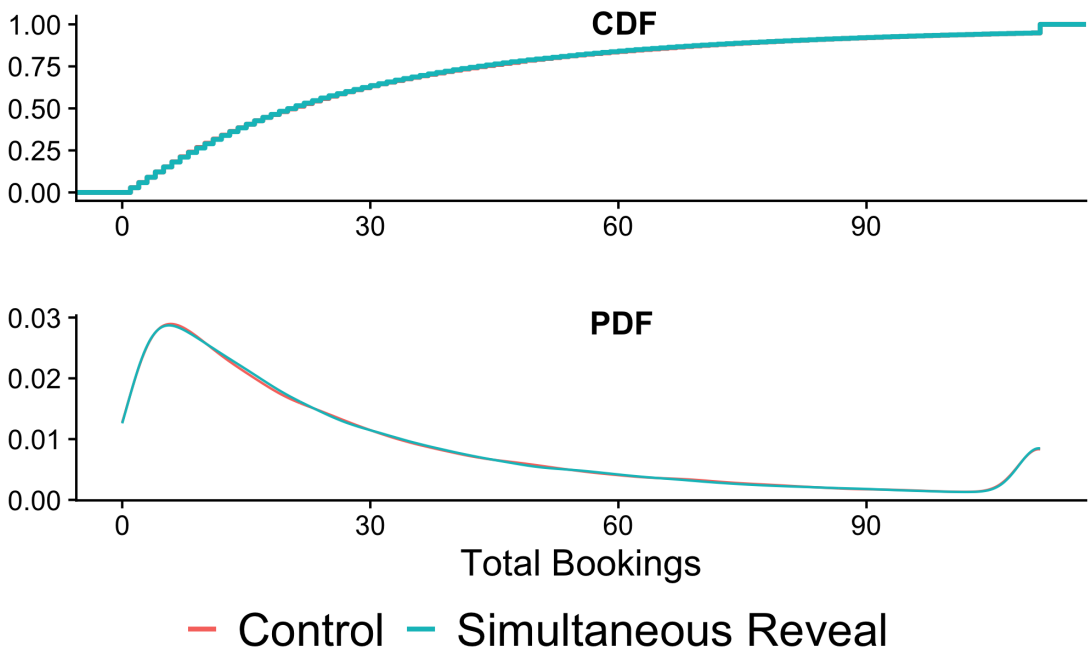
This figure plots the relative likelihood of each review type for compliers (those who review only due to the treatment) relative to the rate for always reviewers (those who would review regardless of treatment). Confidence intervals are computed using the 'basic' method.

Figure A3: Always Reviewer Causal Effects - Monotonicity Robustness



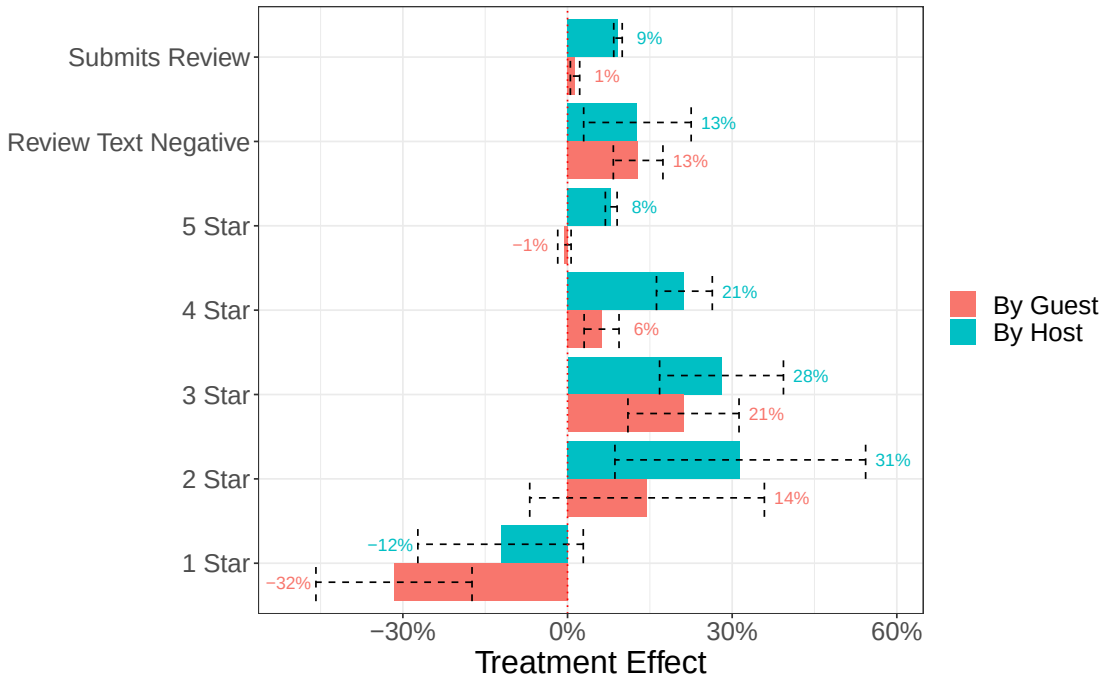
This figure displays the estimated effects of the treatment on always reviewers with the assumption that there are defiers (those who review in the control but not the treatment). We assume that the number of defiers is 33% the number of compliers (those who review in the treatment but not in the control).

Figure A4: Distribution of Bookings by January 1, 2015



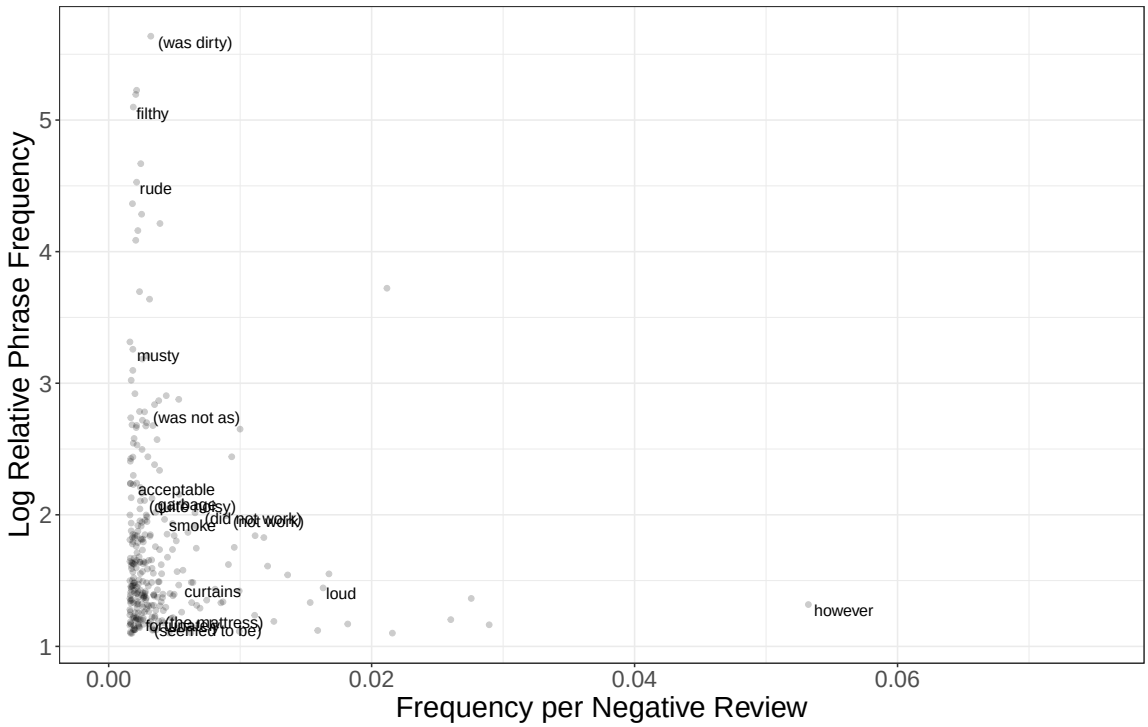
The above figure displays the empirical CDFs and PDFs of total bookings for treated and control listings up to January 1, 2015. We censor the number of bookings at the 95^{th} percentile to make the figure easier to read.

Figure A5: Robustness to Alternative Sample: Treatment Effects



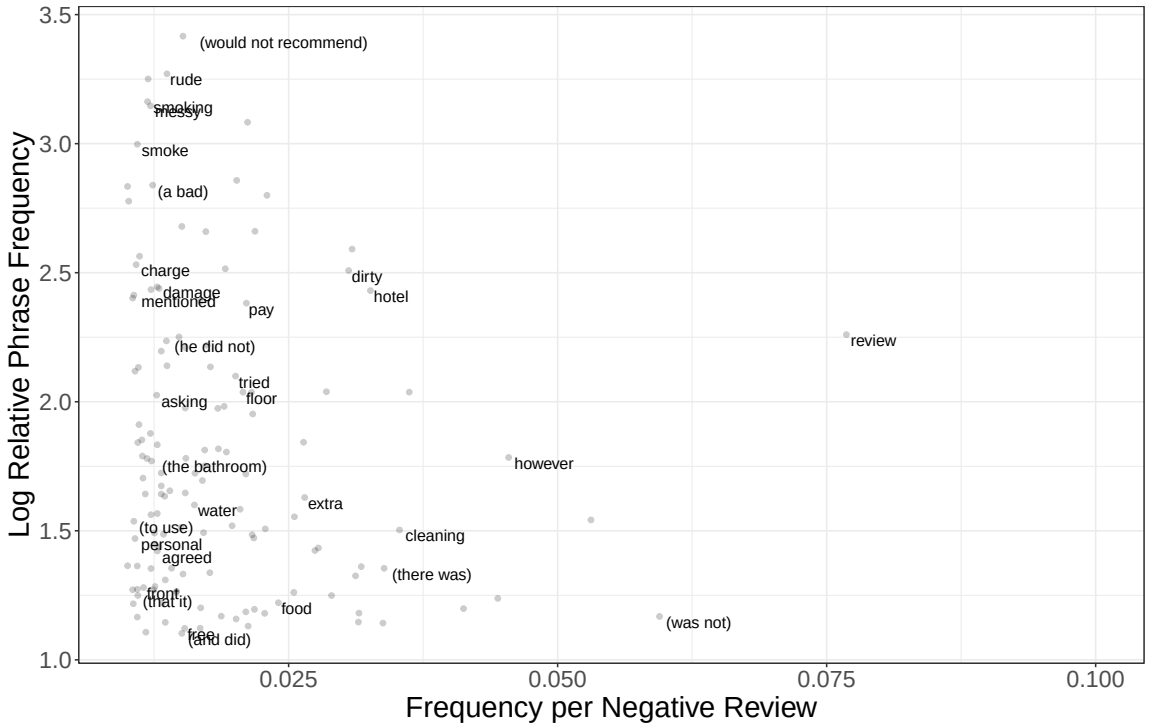
This figure displays the effects on the treatment on reviews by guests and hosts. We measure the percentage effect as the ratio of the absolute treatment effect and the mean in the control. Standard errors used for 95% confidence intervals are computed using the delta method.

Figure A6: Distribution of negative phrases in guest reviews of listings.



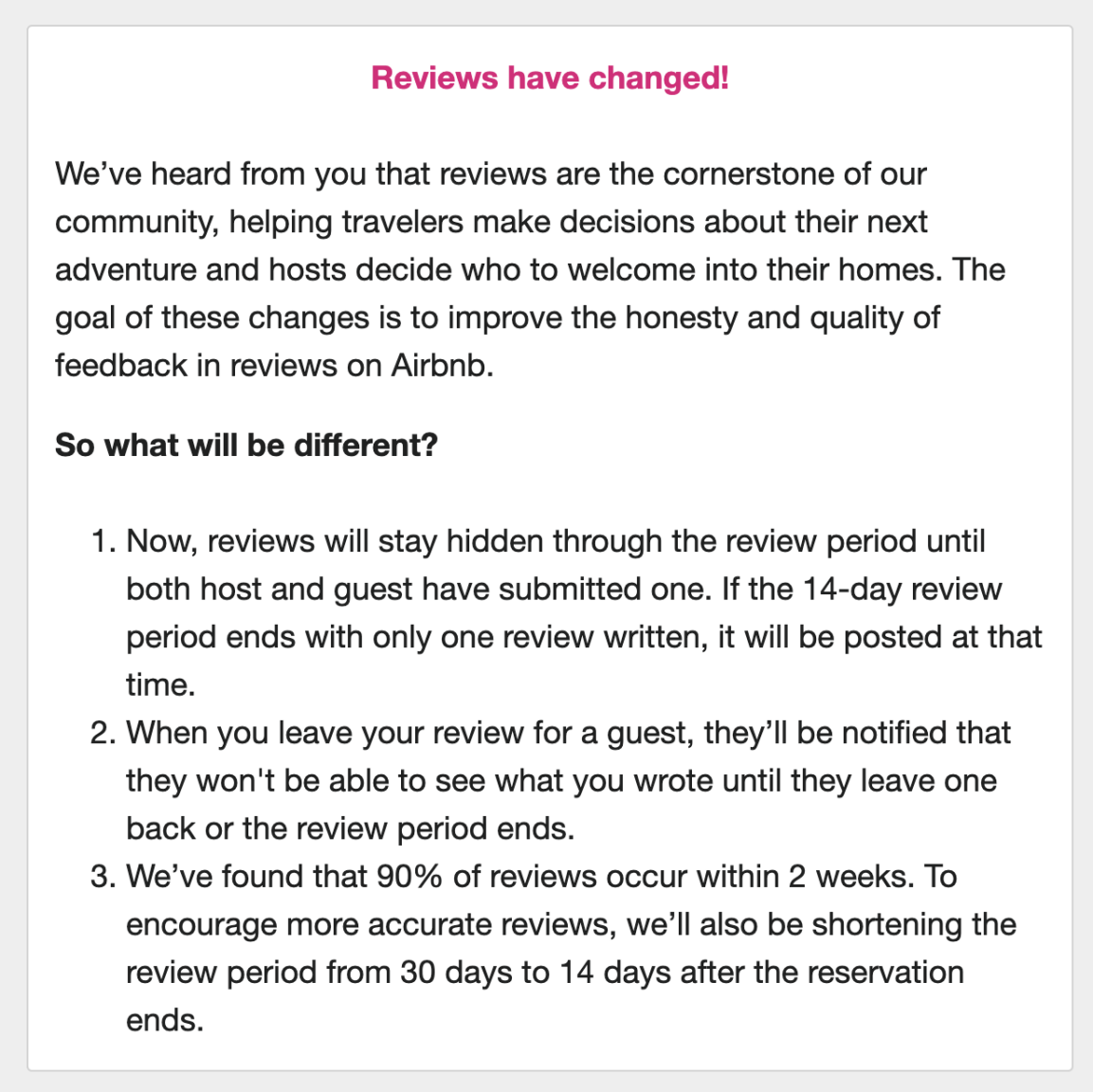
"Relative phrase frequency" refers to the ratio with which the phrase occurs in reviews with a rating of less than five stars.

Figure A7: Distribution of negative phrases in host reviews of guests.



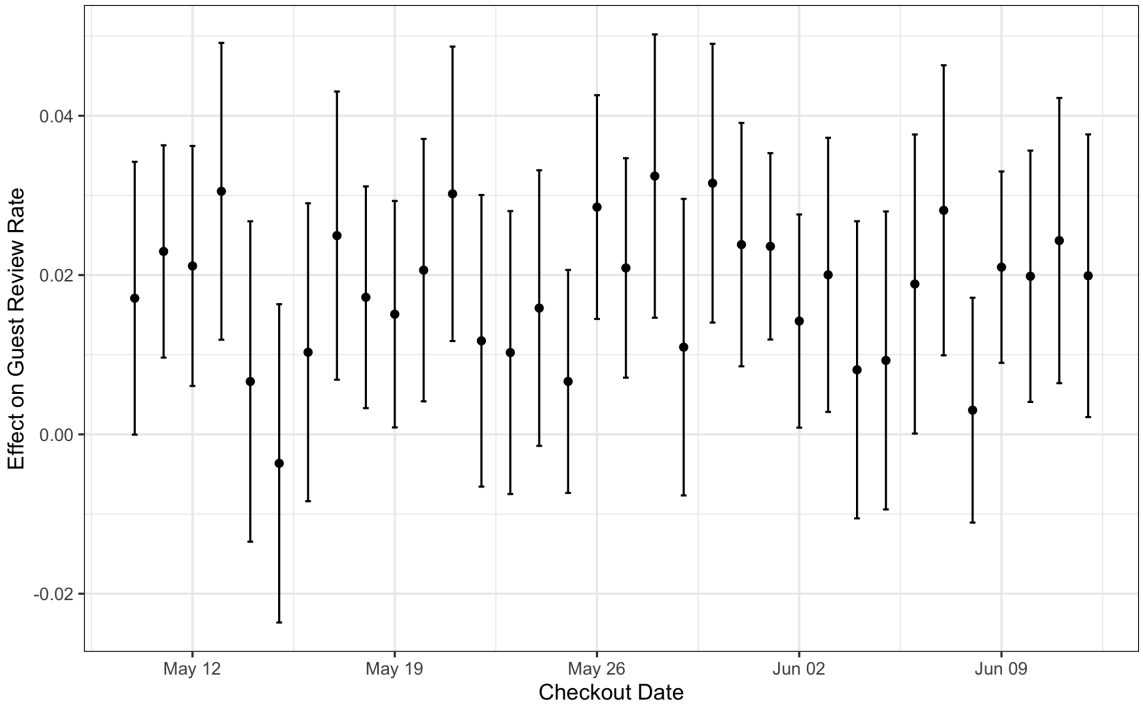
"Relative phrase frequency" refers to the ratio with which the phrase occurs in reviews with a non-recommendation.

Figure A8: Interstitial in Some Host Emails



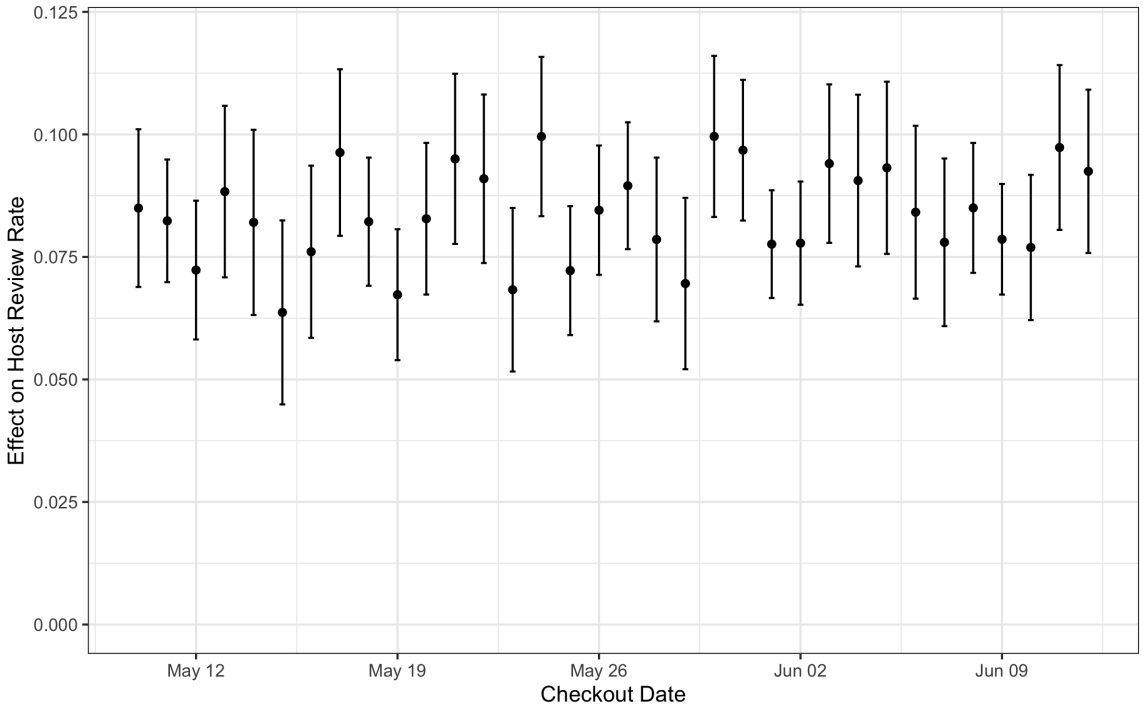
The above figure displays an interstitial inserted into emails received by hosts in the treatment. We are not sure which share of hosts received this interstitial.

Figure A9: Effect of Treatment on Guest Review Rates Over Time



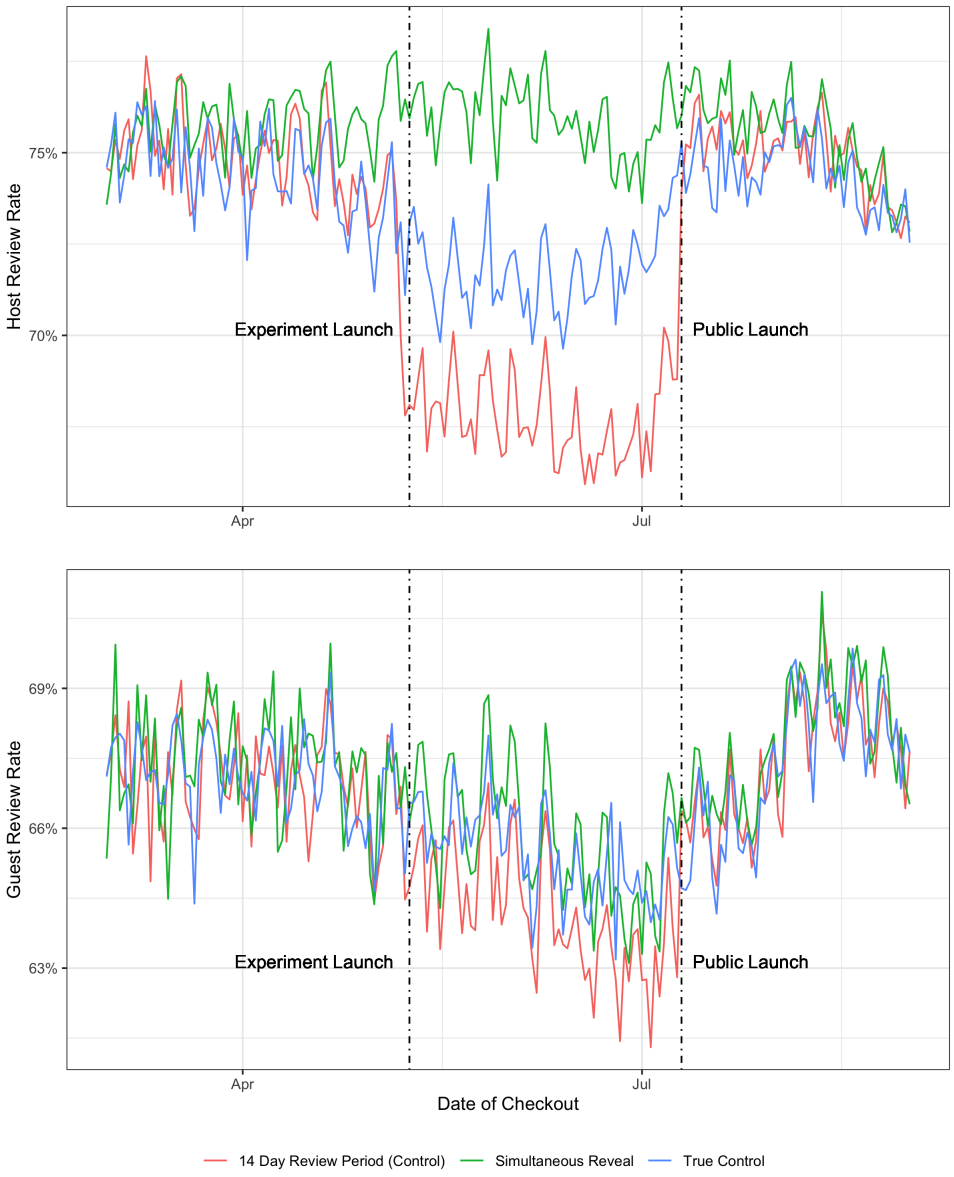
This figure plots the daily treatment effect on guest review rates and the 95% confidence interval. We use all transactions in the sample to increase precision.

Figure A10: Effect of Treatment on Host Review Rates Over Time



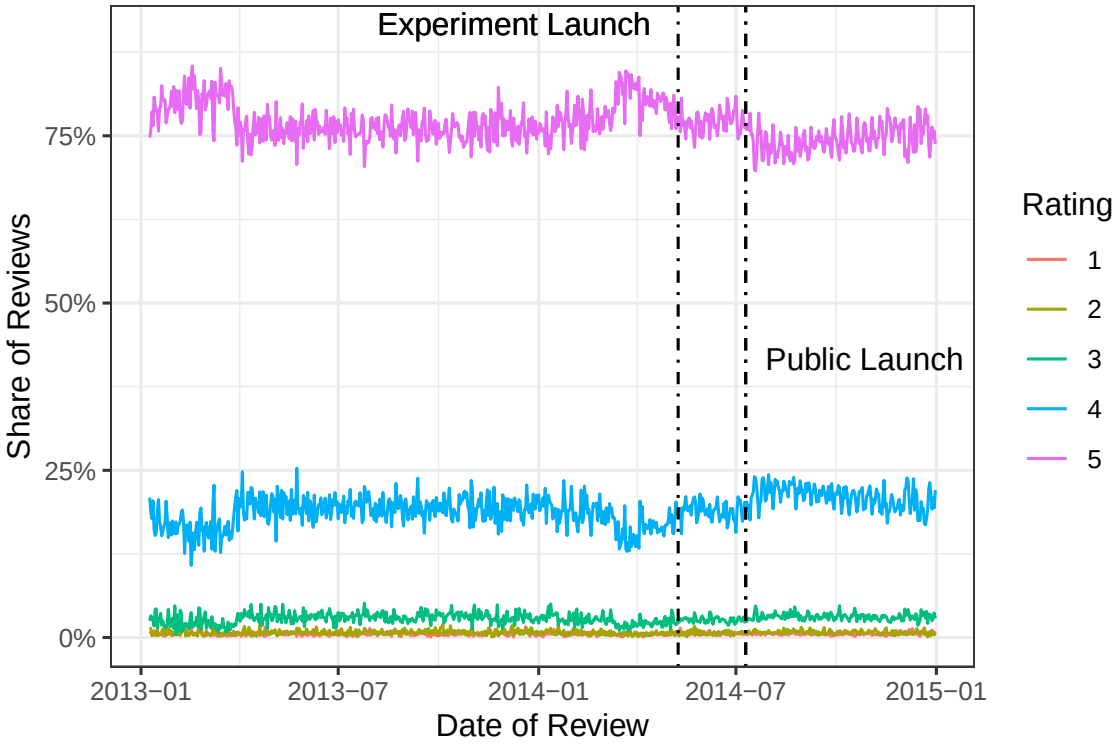
This figure plots the daily treatment effect on guest review rates and the 95% confidence interval. We use all transactions in the sample to increase precision.

Figure A11: Review Rates Over Time



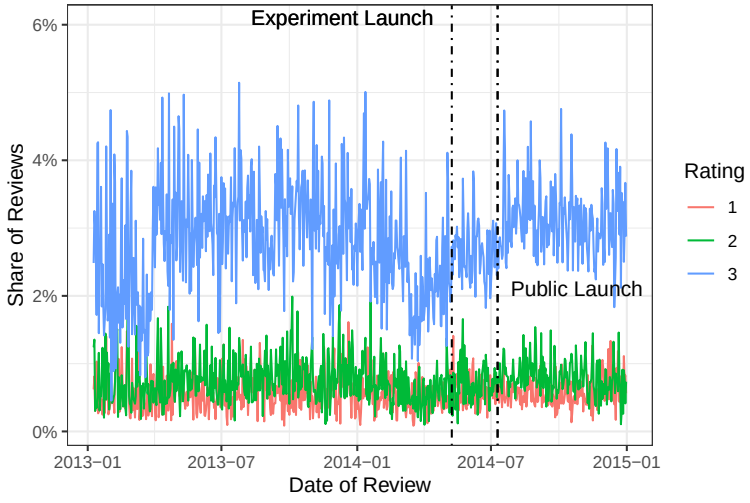
This figure displays the temporal trends of host and guest review rates over time by treatment group. Note that the Simultaneous Reveal Treatment changed the review period to 14 days from 31 days (True Control).

Figure A12: Ratings Over Time



This figure displays the temporal trends of star ratings over time. Because the composition of guests and hosts varies with the growth of the platform, this figure includes only experienced guests reviewing from the domain ("www.airbnb.com") who stayed in a US based listing. The first vertical line demarcates the start of the experiments while the second line demarcates the public launch of the simultaneous reveal system, which placed an additional two-thirds of hosts into the simultaneous reveal treatment.

Figure A13: Ratings Over Time - Low Ratings



This figure displays the temporal trends of star ratings over time. Because the composition of guests and hosts varies with the growth of the platform, this figure includes only experienced guests reviewing from the domain ("www.airbnb.com") who stayed in a US based listing. The first vertical line demarcates the start of the experiments while the second line demarcates the public launch of the simultaneous reveal system, which placed an additional two-thirds of hosts into the simultaneous reveal treatment.