

Marketing Science

Publication details, including instructions for authors and subscription information:
<http://pubsonline.informs.org>

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

Andrey Fradkin, Elena Grewal, David Holtz

To cite this article:

Andrey Fradkin, Elena Grewal, David Holtz (2021) Reciprocity and Unveiling in Two-Sided Reputation Systems: Evidence from an Experiment on Airbnb. *Marketing Science* 40(6):1013-1029. <https://doi.org/10.1287/mksc.2021.1311>

Full terms and conditions of use: <https://pubsonline.informs.org/Publications/Librarians-Portal/PubsOnLine-Terms-and-Conditions>

This article may be used only for the purposes of research, teaching, and/or private study. Commercial use or systematic downloading (by robots or other automatic processes) is prohibited without explicit Publisher approval, unless otherwise noted. For more information, contact permissions@informs.org.

The Publisher does not warrant or guarantee the article's accuracy, completeness, merchantability, fitness for a particular purpose, or non-infringement. Descriptions of, or references to, products or publications, or inclusion of an advertisement in this article, neither constitutes nor implies a guarantee, endorsement, or support of claims made of that product, publication, or service.

Copyright © 2021, INFORMS

Please scroll down for article—it is on subsequent pages



With 12,500 members from nearly 90 countries, INFORMS is the largest international association of operations research (O.R.) and analytics professionals and students. INFORMS provides unique networking and learning opportunities for individual professionals, and organizations of all types and sizes, to better understand and use O.R. and analytics tools and methods to transform strategic visions and achieve better outcomes.

For more information on INFORMS, its publications, membership, or meetings visit <http://www.informs.org>

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

Andrey Fradkin,^{a,b} Elena Grewal,^c David Holtz^{b,d}

^aMarketing, Boston University, Cambridge, Massachusetts 02139; ^bMIT Initiative on the Digital Economy, MIT Sloan School of Management, Cambridge, Massachusetts 02142; ^cYale School of the Environment, New Haven, Connecticut 06511; ^dManagement of Organizations and Entrepreneurship and Innovation, Haas School of Business, University of California, Berkeley, California 94720

Contact: fradkin@bu.edu,  <https://orcid.org/0000-0002-3238-0592> (AF); elena.grewal@airbnb.com (EG); dholtz@mit.edu,

 <https://orcid.org/0000-0002-0896-8628> (DH)

Received: April 20, 2018

Revised: June 5, 2019; March 30, 2020; December 26, 2020; March 31, 2021

Accepted: April 3, 2021

Published Online in Articles in Advance: October 8, 2021

<https://doi.org/10.1287/mksc.2021.1311>

Copyright: © 2021 The Author(s)

Abstract. Reputation systems are used by nearly every digital marketplace, but designs vary and the effects of these designs are not well understood. We use a large-scale experiment on Airbnb to study the causal effects of one particular design choice—the timing with which feedback by one user about another is revealed on the platform. Feedback was hidden until both parties submitted a review in the treatment group and was revealed immediately after submission in the control group. The treatment stimulated more reviewing in total. This is due to users' curiosity about what their counterparty wrote and/or the desire to have feedback visible to other users. We also show that the treatment reduced retaliation and reciprocation in feedback and led to lower ratings as a result. The effects of the policy on feedback did not translate into reduced adverse selection on the platform.

History: Avi Goldfarb served as the senior editor for this article.

 **Open Access Statement:** This work is licensed under a Creative Commons Attribution 4.0 International License. You are free to copy, distribute, transmit and adapt this work, but you must attribute this work as "Marketing Science. Copyright © 2021 The Author(s). <https://doi.org/10.1287/mksc.2021.1311>, used under a Creative Commons Attribution License: <https://creativecommons.org/licenses/by/4.0/>."

Conflict of Interest Statement: Fradkin, Holtz, and Grewal were employed by Airbnb, Inc. for part of the time that this paper was written and have held material financial positions in Airbnb during the life cycle of this paper.

Supplemental Material: Data and the online appendix are available at <https://doi.org/10.1287/mksc.2021.1311>.

Keywords: reputation systems • field experiments • market design • digital platforms • behavioral economics

1. Introduction

Reputation systems are used by nearly every digital marketplace to reduce problems stemming from information asymmetry and moral hazard. They do so by soliciting information about transaction quality and displaying it to other market participants. However, the creation of accurate reviews by market participants is voluntary and costly. As a result, reviews are underprovided in the absence of an appropriate compensation scheme (Avery et al. 1999). This leads to missing information and a variety of biases, which can affect outcomes for both buyers and sellers. For instance, prior work shows that an upwardly biased reputation system can cause buyers to unexpectedly transact with low-quality sellers, which in turn makes them less likely to transact on that platform again in the future (Nosko and Tadelis 2015). These factors make the design of effective reputation systems important for digital platforms.

We study the effects of an experimental change to Airbnb's reputation system. The system is two sided, meaning that the guest and the host each have the opportunity to review one another. In the control group of our experiment, reviews are revealed both to the counterparty and to the public as soon as they are submitted. This leaves open the possibility that the second reviewer reciprocates or retaliates against the first review. Prior research has suggested that this type of behavior occurs on eBay and other platforms with bilateral review systems (Cabral and Hortaçsu 2010, Bolton et al. 2012). Our simultaneous reveal (SR) treatment changes the timing with which a review is revealed to the counterparty and on the platform. In the treatment group, reviews are hidden until both parties have reviewed or until the time to review (14 days) has expired. After reviews are revealed, they cannot be modified, which makes it impossible for users in the treatment group to retaliate against a negative review.

We study the effects of this treatment, first studied in the laboratory by Bolton et al. (2012), on the types of reviews that are submitted and on the subsequent demand for treated sellers.

The treatment reduced the time to review by 17% for guests and 9.9% for hosts. It also increased review rates by 1.7% for guests and 9.8% for hosts. We hypothesize that these effects are largely driven by an explanation that has not previously been documented, *the desire to unveil reviews*. This desire may be caused by curiosity, or by a strategic incentive to have information revealed more quickly to future trading partners. In support of this explanation, we show that whereas the treatment decreased the time from check-out to first review by 9.7%, it decreased the amount of time between the first and second review by much more (35%). The change in review timing is seen for guests and hosts across all levels of experience. In concordance with predictions from prior literature (Bolton et al. 2012), the treatment also changed the types of reviews that were submitted. The ratings in the treatment were more negative on average, but the effects were small—the average guest rating was just 0.25% lower in the treatment. The treatment also decreased the correlation between guest and host ratings by 48%.

Next, we consider whether the lower ratings in the treatment represent more accurate ratings due to the reduction in reciprocity, or whether they are solely due to changes in who reviews. Because the treatment increased review rates, selection effects are likely to be present (Dellarocas and Wood 2007). We use the methodology of principal stratification (Ding and Lu 2017) to show that the treatment changed the reviewing behavior of individuals who would have reviewed regardless of the treatment. This, in addition to our results regarding the correlation of guest and host ratings, provides evidence that the effects on ratings reflect more accurate reviews in the treatment group.

Last, we consider the effects of the treatment on subsequent host outcomes on the platform. If the reputation system became more informative due to simultaneous reveal, then treated sellers, especially those of lower quality, should see less demand or should invest more in quality. We do not detect causal effects of the treatment on subsequent listing demand. We hypothesize that this lack of detectable effect is because of the small overall effect of the treatment on the realized distribution of ratings. We also test for heterogeneous effects across seller types and find no evidence that *ex ante* worse hosts are hurt by the treatment. Our findings contrast with those of Bolton et al. (2012) and Hui et al. (2018), who find that similar changes to reputation systems decreased demand for low-quality sellers. We attribute this contrast to a number of factors, which we discuss in greater detail in Section 2.

The rest of this paper proceeds as follows. In Section 2, we describe the related literature in greater

detail. Next, in Section 3, we discuss the theoretical framework for the study. Section 4 describes the setting of Airbnb. In Section 5, we discuss the experimental design, and in Sections 6 and 7, we discuss treatment effects and evidence regarding the importance of unveiling and reciprocity. Section 8 contains the results of robustness checks, and Section 9 contains results pertaining to the effects of the experiment on adverse selection. Last, we conclude and discuss the implications of our results for reputation system design.

2. Literature Review

We contribute to three related research themes within the study of reputation systems. The first research theme studies why people submit feedback and whether this voluntary process produces bias. The second research theme concerns the effects of reputation system design on submitted reviews and subsequent market outcomes in two-sided markets. The third research theme concerns reciprocity and trust on digital platforms including Airbnb.

Because the majority of reputation systems do not contain a payment scheme, the number, accuracy, and selection of realized reviews is determined by behavioral factors and the details of a particular reputation system. Avery et al. (1999) show that evaluations will be underprovided in equilibrium without an appropriate payment scheme, and Miller et al. (2005) show how to design a scoring system with accurate reporting of feedback in equilibrium. These factors have been shown to matter in practice. Dellarocas and Wood (2007) argue, using data from eBay, that people with worse experiences are less likely to submit feedback. Subsequently, Nosko and Tadelis (2015), Cabral and Li (2012), Lafky (2014), Fradkin et al. (2015), and Brandes et al. (2020) have used experiments with rankings, coupons, and reminders to provide evidence for this hypothesis and the complementary hypothesis that people with more extreme experiences are more likely to review.

There are other reasons that the reviews collected by a reputation system may be biased. Li and Hitt (2008) argue that early buyers may have different preferences than late buyers, which could cause early reviews to be nonrepresentative. Bondi (2019) provides a model and empirical evidence of this phenomenon in the market for books. Filippas et al. (2018) argue that because reviewers may feel bad hurting a counterparty via a negative review, average review scores may inflate over time on platforms.

There have also been a number of studies focused on the effects of different reputation system designs in two-sided markets. On Airbnb and similar markets, there is potential for adverse selection and moral

hazard on both sides of the market. This fact makes it useful to have a two-sided reputation system. However, two-sided reputation systems may also allow for the conditioning of feedback on a counterparty's first rating, which can create biased feedback due to reciprocity and retaliation. Therefore, market designers may face a trade-off between two-sidedness and bias. Three papers (Bolton et al. 2012, Klein et al. 2016, Hui et al. 2018) study these trade-offs.¹

Bolton et al. (2012) use data from several platforms as well as from laboratory experiments to document retaliation in two-sided review systems. They find that when mutually negative feedback occurs, the second review occurs quickly after the first. This is stated as evidence for retaliation. The authors propose a simultaneous reveal system, like the one studied in our paper, and test it in the laboratory. They find that simultaneous reveal decreases review rates, ratings, and the correlation between ratings.

We conduct and analyze the first field experimental test of such a system. We find small effects on ratings, increases in the number of reviews, and decreases in the correlation of ratings. The differences in our results highlight important trade-offs between field and laboratory experiments. On the one hand, laboratory experiments may miss important features of the economic environment of a proposed policy (Levitt and List 2007). On the other hand, as we discuss in Section 8, the experiment we study is not as "clean" as a laboratory experiment, because of the practical considerations involved in running a large-scale experiment with a company. Differences between the laboratory and our field setting include the social nature of the transaction, the underlying distribution of transaction quality, differences in how information was conveyed, the salience of notifications to review, and the incentive to have reviews revealed quickly. In particular, the incentive to have reviews revealed quickly is an important driver of our results, and this factor is not present in the laboratory experiments of Bolton et al. (2012).

Klein et al. (2016) and Hui et al. (2018) study the effects of eBay's change from a two-sided to a (mostly) one-sided reputation system using a before and after observational study. We discuss these papers jointly because they are similar and provide important evidence on the effects of reputation system design. Klein et al. (2016) argue that the main effect of the change was to reduce strategic bias as measured by retaliatory feedback. They then argue that this reduction in bias leads to a decrease in moral hazard, as measured by an increase in realized post-transaction ratings. In contrast, Hui et al. (2018) argue that the improvement in measured buyer satisfaction is due to a reduction in adverse selection; namely, after the change, low-quality sellers are less demanded even if they do not exit the market. Our paper complements these papers

by studying a related policy in a different but equally important market. Furthermore, we use a randomized control trial, which reduces concerns regarding the internal validity of the study. We do not find evidence that adverse selection was substantially reduced by this policy change.

We find that both the distribution of ratings and the rates of reviewing changed due to the simultaneous reveal treatment. In light of this finding, we call for caution in using realized ratings to measure quality. In both Klein et al. (2016) and Hui et al. (2018), quality is primarily measured through changes in realized detailed seller ratings (DSRs). These papers argue that it is unlikely that the switch to a one-sided system affected DSR reviewing behavior, because DSRs are anonymous and displayed only as averages. Airbnb's star ratings are, like eBay DSRs, anonymous and displayed only as averages to hosts during our study period. We find that these star ratings are affected by the simultaneous reveal treatment, even for the first transaction in the experiment for which there is no possibility of a reduction in moral hazard or adverse selection. Therefore, for Airbnb and similar platforms, ratings cannot be used to measure changes in quality without an explicit model of reviewing behavior.

Last, reputation on Airbnb has been the subject of other academic studies. For example, Zervas et al. (2021) compared ratings of the same accommodation on Airbnb and other digital platforms that have one-sided reviews and found that ratings on Airbnb are higher. We show that strategic considerations do not explain these differences, because they have small effects on ratings. An earlier version of this paper, Fradkin et al. (2015) initially presented in 2014, contained many of the results presented in this work and has influenced subsequent research regarding reputation on Airbnb, including Proserpio et al. (2018) and Jaffe et al. (2019). Proserpio et al. (2018) propose that the social aspect of Airbnb transactions may affect realized quality in addition to reviewing behavior, whereas Jaffe et al. (2019) show how transactions with low-quality sellers reduce guests' subsequent usage of the Airbnb platform.

3. Theoretical Framework

The game of reciprocal reviewing with variable review timing has not, to our knowledge, been formalized in the preceding literature. Prior work has instead made informal arguments about the effects of reciprocal feedback, namely, that positive feedback induces positive feedback in response and that negative feedback triggers retaliation. In this section, we add an additional component to the theory of reciprocal reviewing, which we call *the desire to unveil reviews*. This component provides an incentive for agents

to review more often and more quickly in the simultaneous reveal reputation system. In the subsequent empirical sections, we argue that prior theories of reciprocal reviewing are not sufficient by themselves, and that the desire to unveil reviews helps to explain our results.

We begin by summarizing the arguments in the prior literature. Prior work is implicitly based on a utility function consisting of the following terms:

1. An intrinsic cost (or benefit) from leaving a review. This will vary across individuals so that some people dislike reviewing whereas others like it.

2. A disutility from misrepresenting the quality of a transaction in a review. This implies that absent other forces individuals review honestly.

3. For a second review, a positive utility from submitting a review with a reciprocal rating. For example, the second reviewer may feel obliged to leave a positive review once they read the positive review of the first reviewer.

4. A benefit from having more and better reviews. Because the first review affects the rate and type of second reviews, the reviewer rationally takes this into account. The leads to more positive ratings and fewer negative ratings in a first review relative to the situation where just the other components of the utility function are present.

We can now compare the utility of reviewing in the SR versus the review in turn (RIT) reputation systems. We discuss the implications separately for the second reviewer and then for the first reviewer. Throughout the discussion below, we assume that the opportunity to review first or second is exogenous.²

The second reviewer knows that her review will not change anything about the first review. As a result, term 4 drops out under both systems. In the SR system, the second reviewer does not know the content of the first review. As a result, she is not affected by the content of the first review, and term 3 drops out. In contrast, in the RIT system, term 3 remains. Because term 3 increases the utility of reviewing, the utility of reviewing is lower in the SR system and the review rate is predicted to fall. This is the argument made by Bolton et al. (2012) and confirmed in their laboratory experiment.

The first reviewer in the SR system knows that the content of the first review will not affect the content of the second review. As a result, the first reviewer has less reason to review and to do so positively in the SR system than in the RIT system. Consequently, if the only factor driving reviewing behavior were reciprocity, then first reviewer's review rate would be predicted to fall as well.

We now add one more term to the utility function, which corresponds to the desire to unveil reviews. Our theory states that the presence of a "hidden" first

review that can be revealed increases the utility of reviewing. Suppose you are a host who has received an email notifying you of a new guest review. Naturally, you would like to know what the guest said, but you cannot until you review the guest. Furthermore, if you expect the review is positive, you might also want it displayed publicly as quickly as possible, so that you can receive more bookings. This combination of curiosity and strategic behavior motivates you to leave a review right away, rather than wait.

Although it is standard to take strategic considerations into account when studying markets, the role of curiosity has less frequently been considered. The information conveyed in a review is similar to gossip and other social information, which is the topic of much of human conversation and has been shown to trigger curiosity (Dunbar et al. 1997). More generally, curiosity has been shown to strongly affect behavior (see Loewenstein 1994, Silvia 2012). Given the similarities between online reviews and other sources of social information, curiosity should also be present in the setting of reviews.

The key consequence of the desire to unveil reviews is that there will be more reviews and faster reviews in the SR system. In particular, the second review should arrive faster in the SR system than in the RIT system. This prediction is opposite to the prediction yielded by the standard reciprocity motive.

The desire to unveil reviews may also have an effect on the speed of first reviews; namely, users may want to write a quick first review in order to trigger a quick second review. We posit that the effect on first review timing will be smaller than the effect on second review timing, because the existence of a treatment effect for first review timing depends on first reviewers understanding that a quick first review will trigger a quick second review. Given that we had to write this research paper, it is reasonable to assume that many first reviewers have not considered this implication of the SR system.

To conclude the theoretical framework section, we consider the effects of switching to an SR system on ratings. The removal of the ability to condition the second rating on the first in the SR system should result in less inflated ratings and less correlation between reviews (Bolton et al. 2012). The desire to unveil reviews does not change this implication. Even in the SR system, there may still be some reciprocity. For example, the host may give a negative rating in anticipation of a negative rating from the guest. Nonetheless, reviews in the SR system should be less influenced by reciprocity and should be more correlated with the underlying quality of the transaction (Hui et al. 2018, Klein et al. 2016). If reviews are more informative, then worse listings are less likely to be booked, or will be forced to lower prices; that is, SR should reduce

adverse selection. We measure the magnitude of the effects on adverse selection in Section 9.

4. Setting

Airbnb describes itself as a trusted community marketplace for people to list, discover, and book unique accommodations around the world. Airbnb has intermediated over 400 million guest stays since 2008 and lists over five million accommodations. Airbnb has created a market for a previously rare transaction: the rental of an apartment or part of an apartment for a short-term stay by a stranger.

In every transaction, there are two parties: the host, to whom the listing belongs, and the guest, who has booked the listing. After the guest checks out of a listing, there is a period of time (equal to 14 days for the experimental analysis and 30 days for the pre-experimental sample) during which both the guest and host can review each other.³ Both the guest and host are prompted to review via email the day after checkout. The host and guest are also shown reminders to review their transaction partner if they log onto the Airbnb website or open the Airbnb app (and may receive notifications related to reviews). Reminders are also sent when the counterparty submits a review or if the reviewer has not left a review after certain, predetermined lengths of time. Users cannot change their reviews after they have been submitted.

At the time of the simultaneous reveal experiment, Airbnb’s prompt for guest reviews of listings consisted of two pages asking public, private, and anonymous questions (shown in Figure 1). On the first page, guests were asked to leave feedback consisting of

publicly displayed text, a one-to-five-star rating,⁴ and private comments to the host.

The next page asked guests to rate the listing in six specific categories: accuracy of the listing compared with the guest’s expectations, the communication of the host, the cleanliness of the listing, the location of the listing, the value of the listing, and the quality of the amenities provided by the listing. Rounded averages of the ratings were displayed on each listing’s page once there were at least three submitted reviews. The second page also contained a question that asked whether the guest would recommend staying in the listing being reviewed. Overall ratings and review text were required and logged more than 99.9% of the time conditional on a guest review.⁵

The review prompt for host reviews of guests was slightly different. Hosts were asked whether they would recommend the guest (yes/no) and to rate the guest in three specific categories: the communication of the guest, the cleanliness of the guest, and how well the guest respected the house rules set forth by the host. Hosts were not asked to submit an overall star rating. The answers to these questions are not displayed anywhere on the website. Hosts also submitted written reviews that are publicly visible on the guest’s profile page. Finally, the hosts could provide private text feedback about the quality of their hosting experience to the guest and separately to Airbnb.

5. The Simultaneous Reveal Experiment

We now describe the design of the simultaneous reveal experiment and reviewing patterns in the control group. Prior to May 8, 2014, both guests and hosts had 30 days after the checkout date to review each

Figure 1. (Color online) Review Flow on the Website

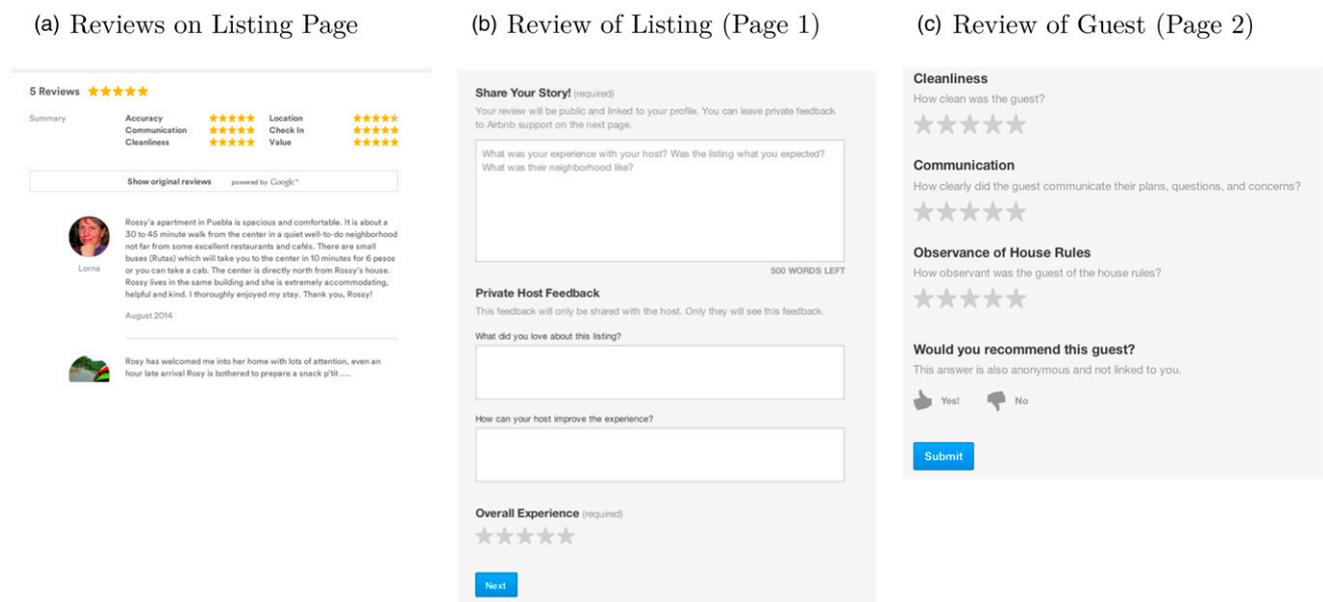
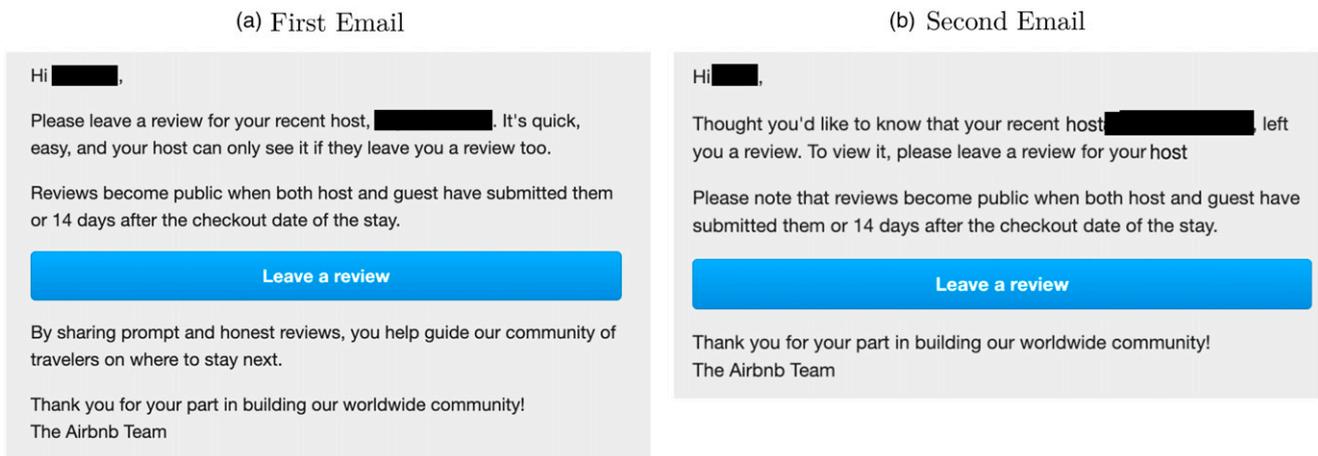


Figure 2. (Color online) Simultaneous Review Email—Guest

other and any submitted review was immediately posted to the website. This allowed for the possibility that the second reviewer could retaliate against or reciprocate the first review. Furthermore, because of this possibility, first reviewers could strategically choose to not review or attempt to induce a reciprocal response by the second reviewer.

The experiment precluded this form of reciprocity by changing the timing with which reviews are publicly revealed on Airbnb. Starting on May 8, 2014, one-third of hosts were assigned to a treatment in which reviews were hidden until either both guest and host submitted a review or 14 days had expired. Another third of hosts were assigned to a control group where reviews were revealed as soon as they were submitted and there was a 14-day review period.⁶ Reviews were solicited via email and app within a day of the guest's checkout. An email was also sent when a counterparty submitted a review. Last, a reminder email was sent close to the end of the review period.

Users in the treatment received different review-related emails from users in the control. Figures 2 and 3 show the emails received by guests upon the end of their stay and when the counterparty left a review first.

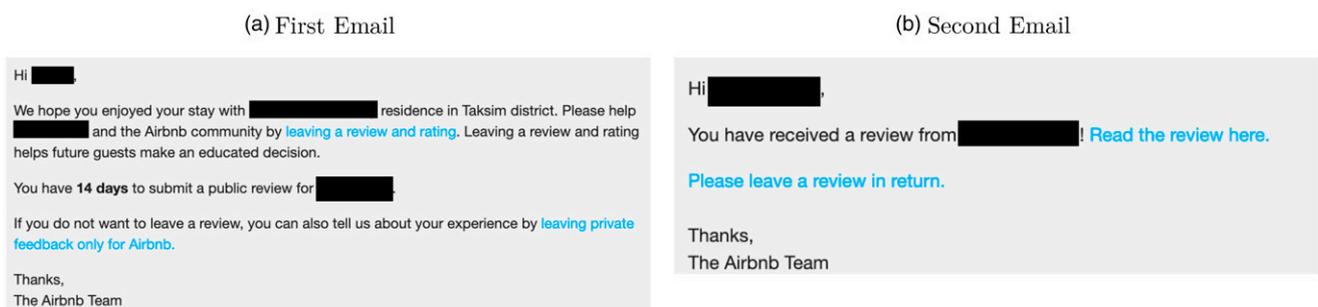
Figure 3. (Color online) Control Email—Guest

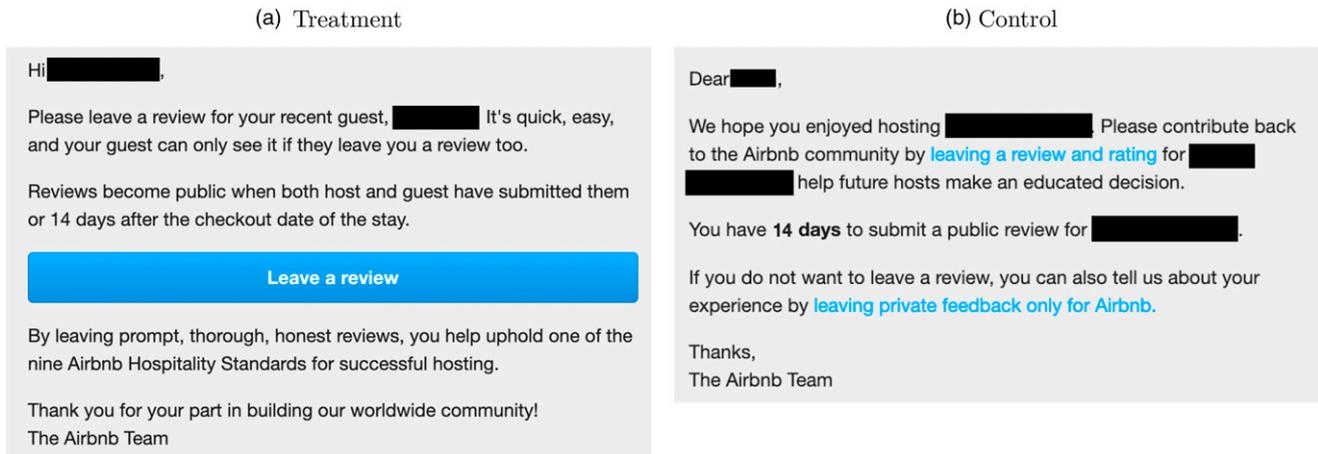
Figure 4 shows the analogous first emails for hosts. During the simultaneous reveal experiment, Airbnb was also making unrelated changes to the content of review-related emails. In Online Appendix F, we discuss the potential impact of these changes on our results. Finally, both guests and hosts received a prominent notification before starting a review (Figure 5).

5.1. Description of Reviewing Behavior in the Control Group

Below, we describe reviewing behavior in the 14-day control. Throughout Sections 5, 6, and 7, we focus on the first transaction observed for each host either in the treatment or in the control.⁷ We later turn to the effects of the experiment on subsequent reviews and stays.

Our baseline sample consists of 119,789 transactions starting with checkout dates on May 10, 2014, and ending with checkout dates on June 12, 2014.⁸ On average, users review frequently and positively, with hosts reviewing more positively and faster than guests. In the control group, 68% of trips result in a guest review, and 72% result in a host review. Reviews are typically submitted within a few days of the

Figure 4. (Color online) Host First Emails



checkout, with hosts taking an average of 3.8 days to leave a review and guests taking an average of 4.7 days. The average time between a first and a second review in the control group is 3 days. This is an important statistic for testing the desire to unveil reviews, and we will return to it in Section 6.

Reviews are mostly positive. Conditional on a review, 74% of guests leave a five-star overall rating, and 48% of guests submit fives for all of the category ratings. Figure 6(a) displays the distributions of ratings for reviews by guests and hosts. Both distributions are skewed toward the right, with the majority of ratings being four and five stars. Host reviews are even more positive than guest reviews, with 86% of host reviews containing five-star ratings for all categories.

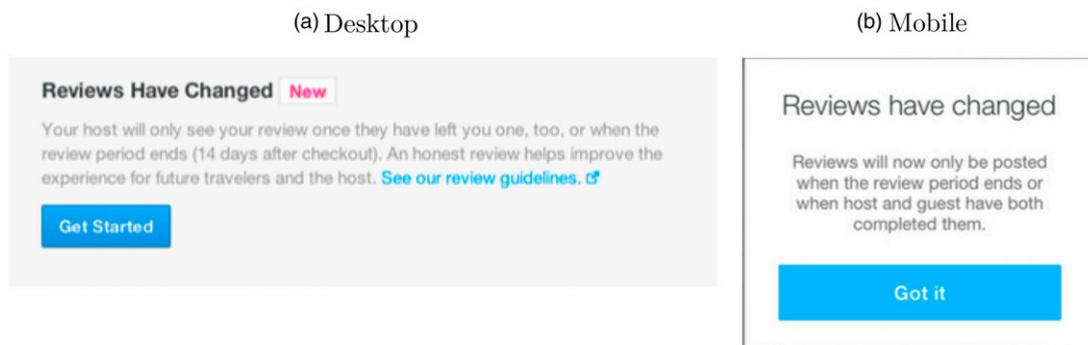
Text comprises another important part of the review that we incorporate into our analysis. We trained a regularized logistic regression model on pre-experiment data to classify the sentiment of reviews and to determine the words and phrases associated

with negative reviews. A discussion of the training procedure can be found in Online Appendix B.

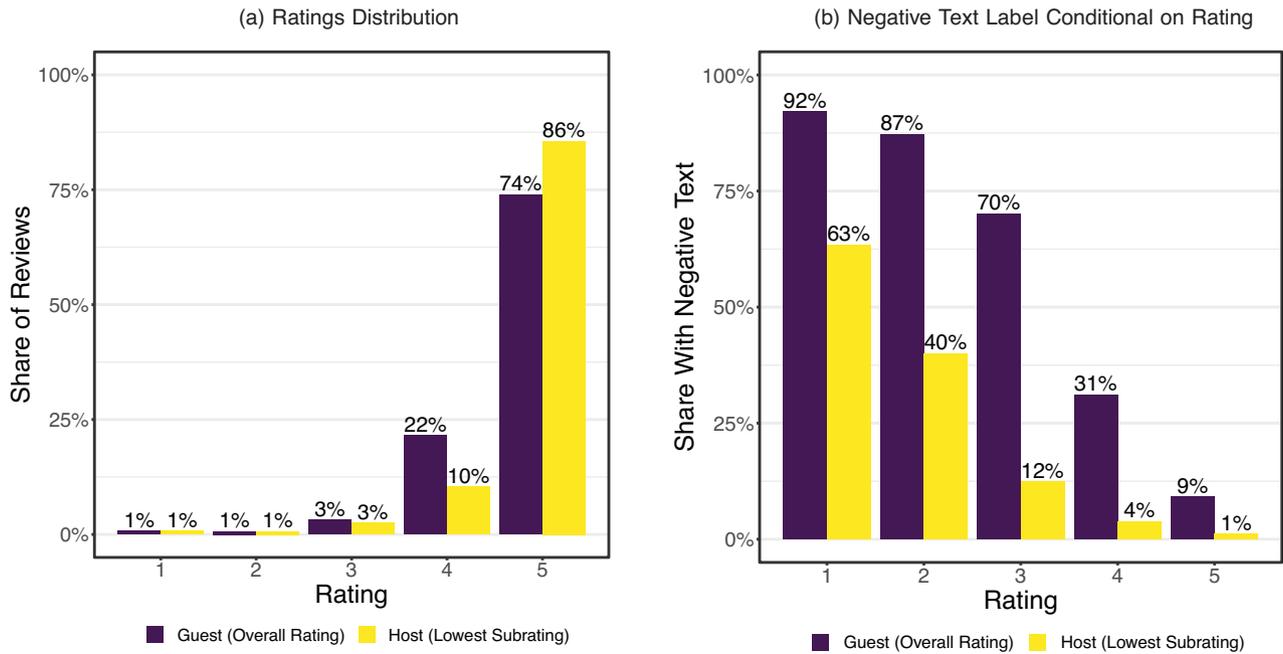
In Figure 6(b), we show the share of negatively labeled text reviews by star rating in the control group. Low star ratings by guests are typically but not always associated with negative text. Ninety percent of one- and two-star reviews by guests are classified as negative, whereas three-star reviews have text that is classified as negative 70% of the time. Hosts are less willing to leave negative text even when they leave a low category rating for the guest.

With regard to more positive reviews, negative text is less prevalent but still exists. Guests write negatively classified text 31% of the time for four-star reviews and 9.2% of the time for five-star reviews. This may be due to the desire for guests to explain shortcomings, even if they had a good experience. Another explanation, especially relevant to five star reviews, is measurement error in our text classification procedure.

Figure 5. (Color online) Simultaneous Reveal Notification



Notes. The figure displays the notifications shown to guests prior to seeing the review form. For hosts, the desktop notification had the word "host" replaced with the word "guest."

Figure 6. (Color online) Rating Distributions

Notes. The left panel displays the distribution of submitted overall ratings by guests and lowest category ratings by hosts in the control group of the simultaneous reveal experiment. The right panel displays the prevalence of negative text review as predicted by a regularized logistic regression conditional on rating.

6. The Desire to Unveil Reviews

In this section, we provide experimental evidence in support of users' desire to unveil reviews. As discussed in Section 3, if reviewing is driven by this desire, then the SR treatment should increase review rates and the speed of reviews, particularly following a first review. In contrast, if the main effect of SR is to reduce reciprocity, then we would expect review rates to fall.

Table 1 shows the control and treatment means as well as the treatment effect for review timing-related variables. Review rates increase by 1.7% for the guest and by 10% for the host. Importantly for the unveiling explanation, the number of days between reviews falls by 35%. This is much larger than the fall in the overall time to review (17% for guests and 9.7% for hosts).

Table 1. Summary Statistics

	Control mean	Treatment mean	Effect
<i>Submits Review</i> (guest)	0.68	0.69	0.01***
<i>Submits Review</i> (host)	0.72	0.79	0.07***
<i>Days to Review</i> (guest)	4.70	3.89	-0.81***
<i>Days to Review</i> (host)	3.80	3.42	-0.37***
<i>Days Between Reviews</i>	3.05	1.98	-1.07***
<i>Days to First Review</i>	3.33	3.01	-0.32***

Notes. 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.

*** $p < 0.01$.

The large drop in the time between reviews suggests that the first review triggers a much faster second review in the SR treatment, as predicted by the desire to unveil reviews. We now test for this formally by modeling the time to review as a duration. In particular, we expect that in both the treatment and the control, a first review increases the hazard of the second review. This occurs because the first review automatically triggers an email sent to the counterparty, which reminds the other user to review. Furthermore, we predict that the first review has a larger effect on second review hazard in the SR treatment, because the second reviewer wants to reveal the review.

Our empirical specification is displayed below and represents the canonical Cox proportional hazards model:

$$\lambda_i(t|\mathbf{x}_i) = \lambda_0(t)\exp\{\mathbf{x}_i'\boldsymbol{\beta}\}. \quad (1)$$

In the above equation, $\lambda_i(t)$ is the hazard rate of reviewing for individual i at time t . Our covariates, \mathbf{x}_i , include an indicator for the SR treatment, an indicator for whether the time is after the counterparty (guest or host) has reviewed, and an interaction between treatment and being the first review. We are interested in both the baseline effect of the treatment on reviewing and the interaction term.

Table 2 displays estimates of Cox proportional hazard models of review hazards for guests (columns (1) and (2)) and hosts (columns (3) and (4)). We find that

Table 2. Speed of Review Effects

	Dependent variable			
	Relative Hazard of Review			
	Guest		Host	
	(1)	(2)	(3)	(4)
<i>Treatment</i>	1.080	1.014	1.230	1.108
	$t = 11.003^{***}$	$t = 1.450$	$t = 31.003^{***}$	$t = 12.560^{***}$
<i>After Host</i>		2.470		
		$t = 86.862^{***}$		
<i>Treat × After Host</i>		1.123		
		$t = 8.276^{***}$		
<i>After Guest</i>				1.989
				$t = 63.448^{***}$
<i>Treat × After Guest</i>				1.554
				$t = 30.948^{***}$
<i>Number of Events</i>	82,055	82,055	90,034	90,034
<i>R²</i>	0.001	0.084	0.006	0.089

Notes. This table displays the relative hazard estimated from Cox proportional hazard regressions where the outcome is whether a review is submitted by the guest (columns (1) and (2)) or the host (columns (3) and (4)). The terms *After Host* and *After Guest* refer to an indicator for whether the time was after the submission of the review by the counterparty.

*** $p < 0.01$.

the treatment increases the overall hazard of reviews for guests by 8% (column (1)) and for hosts by 23% (column (3)), which is consistent with the fact that the desire to unveil reviews causes faster reviews.

A key prediction of our theory is that the SR treatment should cause an especially large effect on the speed of the second review relative to the first. This is because the second review instantly reveals the first review. To test for this, we interact the treatment with whether the counterparty has already reviewed (columns (2) and (4)).⁹

We find that a first review increases the hazard of a second review in both the treatment and the control.¹⁰ The interaction effect between the treatment and a first review is 12% for guest second reviews of hosts and 55% for host second reviews of guests. Both of these interaction effects are statistically significant.

For both guests and hosts, we find that the effect of the treatment on reviewing is mostly explained by this interaction. To see this, compare the coefficient on the treatment in columns (1) and (2). It falls from 1.08 to 1.01, meaning that the hazard of guest reviews does not increase in the treatment until a host leaves a review. Similarly, when comparing columns (3) and (4), the baseline effect of the treatment on the hazard rate falls from 1.23 to 1.11, meaning that the treatment mostly increases host reviewing through the increased speed of a second review.

The above evidence is consistent with a large effect of the desire to unveil reviews. Not only do review rates increase, but the hazard model shows that faster second reviews after an initial first review explain most of the total effect. We conclude that the desire to

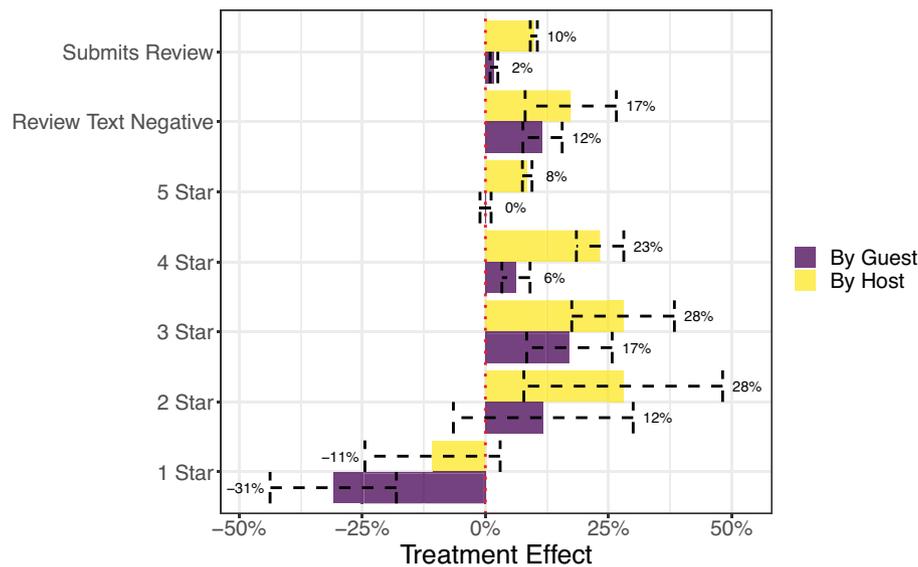
unveil reviews is salient when the second review immediately reveals the contents of the first review.

7. Reciprocity and Its Effects on Ratings

The above section demonstrated that, contrary to the predictions of a model with only reciprocity, the simultaneous reveal treatment caused review rates to increase. This means that the desire to unveil reviews was more influential than reciprocity in determining review rates. We now show that the treatment effects on ratings are consistent with a decrease in reciprocity in the SR treatment.

Recall that because SR eliminates the ability to reciprocate the rating of a first review, ratings should decrease and should be less correlated between the guest and host. Figure 7 displays the treatment effects split by the star rating submitted (we do not condition on whether a review is submitted). Consistent with the first prediction, we see a pronounced increase in reviews with two-to-four-star ratings. On the other hand, reviews with five-star ratings do not increase for guests and increase to a much lesser extent for hosts.

Figure 7 also documents a fall in one-star ratings in the treatment. We posit that this effect is also due to reduced reciprocity; namely, when negative ratings occur in the control group, they often trigger one-star retaliatory reviews. Because the SR treatment prevents guests from seeing the first review content, we observe fewer one-star ratings. One supporting piece of evidence for this explanation is that the fall in one-star ratings by guests is particularly large (66%) for cases

Figure 7. (Color online) Effects of Experiment on Reviews

Notes. This figure displays the percentage change (relative to the mean in the control group) in reviews of a given type due to the treatment. The standard errors used for the 95% confidence intervals are calculated using the delta method. Transactions with no label, such as when there is no review, are treated as zeros for the purpose of this calculation.

in which the first review contains negative text. We see a similar pattern for host reviews after a negative guest review. That said, one-star reviews are rare; there are just 290 one-star reviews in the control group and 201 one-star reviews in the treatment group.

Next, we test the prediction that review content between guests and hosts should be less correlated as a result of the SR treatment. We measure the correlation in reviews across two different measures: the labeled review text and the lowest rating (including subratings). Across both measures, we find large and statistically significant decreases in the correlation of ratings. The correlation of positive text fell by 50% (standard error of 6.7%), and the correlation of ratings fell by 48% (standard error of 4.4%).

In summary, the changes in the observed ratings and the fall in the correlation between guest and host ratings are consistent with the simultaneous reveal treatment reducing reciprocity. In the next section, we consider alternative explanations for our empirical findings, including whether the effects on ratings are caused by changes in who reviews, rather than reductions in reciprocal behavior among those who do review.

8. Alternative Explanations of Experimental Effects

In this section, we consider alternative explanations for the treatment effects of the simultaneous reveal policy. We focus on two main threats to our interpretation. First, it could be the case that increases in

review rates are caused by unintended changes in Airbnb's review solicitation emails. Second, it could be that changes in the submitted ratings are caused by changes in who reviews, rather than in changes to how people review. We discuss both of these threats below and relegate additional robustness concerns to Online Appendix F.

8.1. Do Unintended Changes in the Email Explain Increases in Review Rates?

Recall that the emails in Figure 2 and Figure 3 differed not only in the information that they convey, but also in the size of the "Leave a review" button and the specific email text (i.e., "Thank you for your part in building our worldwide community!"). It could be the case that these confounding changes—and not reciprocity or the review unveiling explanation—explain the treatment effects we observe. Below, we argue that these changes are unlikely to explain the treatment effects, given that the design changes are more pronounced in the first email, but we find larger effects for the second email.

Consider the fact that the large blue button is present for both the first treatment email, which is sent immediately after checkout, and the second treatment email, which is sent after the first review has been left. If the button increased review rates, this increase would manifest for reviews occurring both after the first email and after the second email. Furthermore, only the first email has text asking the reviewer to be "prompt and honest." If this text increased review

rates, it would only have increased the rate of first reviews, not second reviews. Combining these two hypothesized effects, we would expect design changes to the review emails to have a larger effect on the rate of first reviews than on the rate of second reviews. We instead observe that the effects of the treatment are largest for the reviews submitted after the second email (Table 2). In fact, our hazard models show that for guests, the treatment effect on review rates shows up only after the host has submitted a review.

A related point is that there are other ways in which users may learn about the treatment policy, and these are not affected by the confounding email text. Users are alerted to the new policy not only in the review email, but also during the review flow. Anyone logged in on Airbnb.com or on the app is also shown an alert asking them to submit a review. In summary, although we acknowledge that confounding changes to the email text may have effects, we believe they are unlikely to explain the increased review rates we observe.

Last, there was some variation in the email text sent to the treatment group over the course of the experiment. We believe these variations in the email text do not undermine our tests of the desire to unveil reviews, and we provide further evidence for this argument in Online Appendix F.

8.2. Does Selection into Who Reviews Explain the Fall in Average Ratings?

We now consider an alternative explanation for the observed changes to ratings in the treatment. Dellarocas and Wood (2007), Fradkin et al. (2015), and Brandes et al. (2020) all argue that who selects into reviewing affects rating distributions. Because the simultaneous reveal treatment increased review rates, it could be the case that changes in ratings are explained by a change in the composition of reviewers, rather than a fall in reciprocity as argued in Section 7. We use the methodology of principal stratification (Frangakis and Rubin 2002, Ding and Lu 2017) to show that the observed changes in ratings are not caused solely by changes in the selection of reviewers.

Principal stratification is a procedure for identifying the effects of a treatment for latent subgroups in the experimental sample. The effects on these subgroups provide insight into the causal mechanisms underlying the overall treatment effects. In our setting, we posit that there are three latent types of individuals:

Always reviewers. These individuals review regardless of whether they are in SR or the control.

Compliers. These individuals are induced to review by the SR treatment and would not review if they were in the control condition.

Never reviewers. These individuals never review.

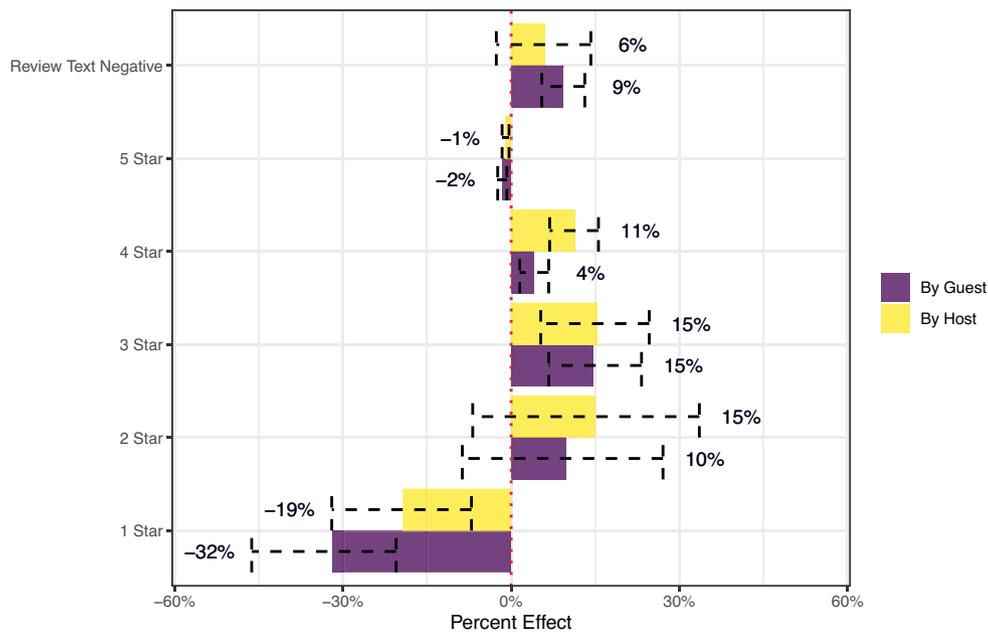
Any effect of the treatment on always reviewers is, by definition, free from selection. The method of principal stratification by principal scores (Ding and Lu 2017) allows us to estimate this treatment effect. The key assumption required for implementing principal stratification is called weak general principal ignorability. It states that the expected outcome, conditional on submitting a review, is independent of latent strata (complier and always reviewer) when controlling for covariates.¹¹ This is a strong condition, but is made more plausible by the availability of pretreatment covariates such as historical ratings by guests and hosts, as well as trip characteristics including whether there were customer service complaints.

The procedure is conducted in several steps. We first use a logistic regression trained on data from the control group to predict the choice of whether to review as a function of user- and trip-level covariates. Similarly, we use a logistic regression in the treatment group to predict the decision to not review using the same covariates. Once we have these probabilities, we can calculate the probability that (conditional on covariates) a user is a never reviewer, always reviewer, or complier. Finally, we can use a weighting procedure to calculate the stratum-specific causal effects. We discuss the details of this procedure in Online Appendix E.

To evaluate the fit of our predictive models, we consider our ability to predict reviewing behavior out-of-sample. We use a 10-fold cross-validation procedure. This procedure produces out-of-sample predictions that we use to calculate the area under the curve (AUC) of the receiver operating characteristic and generate a calibration plot for these predictions. For host reviews of guests, we achieve an AUC of 0.74, whereas for guest reviews of hosts, we achieve a lower AUC of 0.64. Both of these AUC measurements are better than the null of no predictive power. The predictions are also well calibrated (Figure A1 in the online appendix).

Using the principal stratification approach, we find that the treatment *does* change reviewing behavior for the always reviewers—those individuals who would review regardless of treatment status. Figure 8 displays the causal effects for this set of users. We see a pattern of treatment effects consistent with our baseline results. The always reviewers are caused to submit more two-to-four-star ratings relative to one- or five-star ratings as a result of the treatment, and to leave more negative text. In other words, the treatment not only changed which Airbnb users left reviews, but also how Airbnb users reviewed their counterparties conditional on leaving feedback.¹²

One concern about the principal stratification procedure is that it assumes monotonicity, which may be violated if the absence of reciprocity in the treatment causes some individuals to not review. We follow

Figure 8. (Color online) Always Reviewer Causal Effects

Notes. This figure displays the percentage change (relative to the mean in the control) in reviews of a given type due to the treatment. The standard errors used for the 95% confidence intervals are calculated using the percentile bootstrap method. Transactions with no label, such as when there is no review, are treated as zeros for the purpose of this calculation.

Ding and Lu (2017) in testing for the robustness of our results to violations of the monotonicity assumption. We call individuals who would review in the control but not in the treatment defiers, following the standard terminology in settings of experimental noncompliance. We assume that that number of defiers is 33% of the number of compliers and recompute the always reviewer causal effects (Figure A3 in the online appendix). We find very similar results, showing that modest violations of monotonicity do not overturn our findings.

In summary, we have shown that the effects of the simultaneous reveal on ratings are not caused by changes in who selects into reviewing. Instead, they are caused by changes in the ability of reviewers to condition ratings on the content of the first review.

9. Effects on Adverse Selection

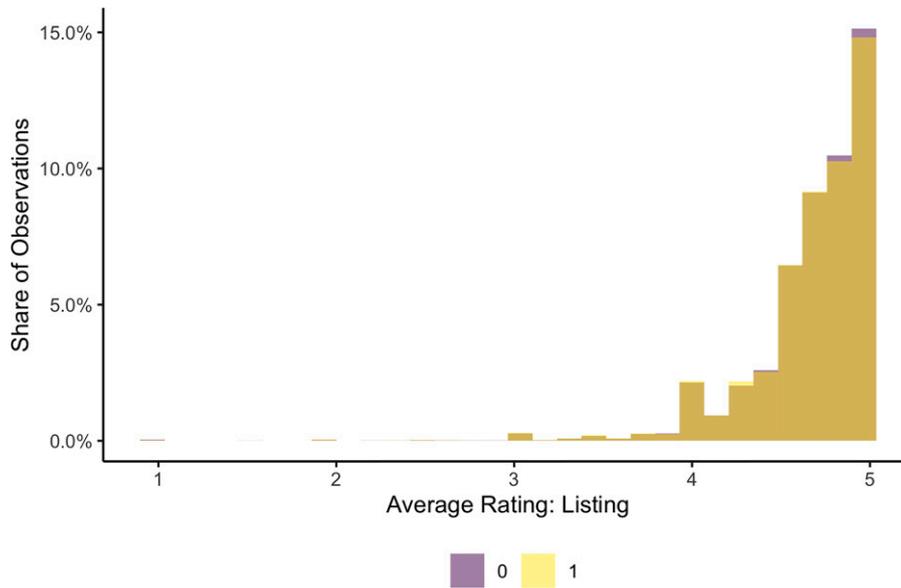
We now discuss the effects of the treatment on the selection of transacting users. If the treatment had its intended effect, then transactions with low-quality users should become less likely in the treatment and transactions with high-quality users should become more likely (Airbnb 2014). Prior observational and laboratory work studying bilateral reputation systems has argued that removing retaliation and reciprocity reduces adverse selection for sellers (Hui et al. 2018). We use our experiment to study the effects of simultaneous reveal on adverse selection in Airbnb and find precisely estimated null effects.¹³

We begin by describing the ways in which simultaneous reveal may affect adverse selection. First, simultaneous reveal reviews were less influenced by reciprocity, which should in theory make them more reflective of user experiences. This more accurate information should create better (although possibly fewer) matches as it redistributes demand from worse listings to better listings.¹⁴ However, the simultaneous reveal policy does not just cause an increase in review accuracy—it also increases the speed and total number of reviews due to the desire to unveil reviews. Because induced reviews are typically positive, this may cause an increase in demand for the treated listings, which are more likely to be rated.

One way to measure the potential impact of the policy is to consider the distribution of average ratings for listings in the treatment and control groups after the first review. Because hosts have already accumulated many reviews, the initial effect of the policy on average ratings at the listing level is small. We plot the difference in realized average ratings for the treatment and control groups in Figure 9. We find small differences, with the control group having slightly more listings with an average rating close to five stars.

Next, we measure the effects of the treatment on listing outcomes. We focus on two kinds of outcomes. The first are measured exclusively in the experimental period. During the experimental period, the differences between reviews in the treatment groups should be most pronounced. However, there will have been

Figure 9. (Color online) Distribution of Average Ratings at a Listing Level



Note. This figure displays the distribution of the average rating across reviews within a listing following the first transaction in the experiment.

less time for those reviews to affect subsequent guest and host behavior.

Table 3 shows precisely estimated zeros for the log of nights in the experimental period and log of average booked price per night in the experimental period. The estimates for the log of revenue are less precise but are still not statistically distinguishable from zero.

We then look at outcomes through the end of 2014. Note that because the treatment was launched platform-wide in July 2014, both treatment groups were partially treated using this outcome metric. We find precisely estimated zeros on the log of bookings through 2015 and whether the listing is active in 2015. In summary, the exposure of listings to the simultaneous reveal treatment does not affect aggregate demand.

As discussed above, we also predict that worse-quality listings should receive less demand than high-

quality listings as a result of the treatment.¹⁵ Such a decrease in demand for ex ante worse listings would represent a reduction in adverse selection. We propose two proxies for listing quality that are unaffected by the treatment and use these to test for heterogeneous treatment effects.

Table 4 displays the specifications that interact the treatment with measures of listing quality. We add two interaction variables. The first of these is the ratio of five star ratings to total transactions occurring prior to the experiment. We call this the effective positive percentage (EPP) as in Nosko and Tadelis (2015), who argue that this is a good proxy for quality. We also add an indicator for whether we can measure the EPP because it is undefined when there are no prior transactions. As intended, a higher EPP is associated with better subsequent listing outcomes even in the control,

Table 3. Treatment Effects on Listing Outcomes

	Dependent variable				
	Log(Nights in Exp.) (1)	Log(Price in Exp.) (2)	Log(Rev. in Exp.) (3)	Log(Bookings by 2015) (4)	Active in 2015 (5)
Treatment	-0.010 (0.006)	-0.005 (0.005)	-0.025 (0.019)	0.002 (0.004)	-0.003 (0.002)
Controls included	Yes	Yes	Yes	Yes	Yes
Observations	119,550	73,234	119,550	119,550	119,550
R ²	0.262	0.411	0.219	0.631	0.078

Notes. This table displays the treatment effects on listing outcomes after the first transaction in the experiment. Controls are included for greater-than-median EPP, whether the EPP is calculable, log of prior bookings, log of the first price, and whether the guest submitted a customer service complaint. Columns with “in Exp.” in the dependent variable refer to outcomes calculated only through June 12, 2014, the end of the experimental period. In column 3, “Rev.” refers to the revenue of the listing during the experimental period. There are fewer observations for the price variable, because we cannot measure transaction prices for hosts who did not transact after the initial transaction in the experiment.

Table 4. Tests of Adverse Selection

	Dependent variable				
	Log(Nights in Exp.) (1)	Log(Price in Exp.) (2)	Log(Rev. in Exp.) (3)	Log(Bookings by 2015) (4)	Active in 2015 (5)
<i>Treatment</i>	−0.009 (0.009)	−0.007 (0.007)	−0.023 (0.029)	0.005 (0.006)	−0.001 (0.004)
>Median EPP	0.067*** (0.010)	0.049*** (0.007)	0.222*** (0.030)	0.054*** (0.006)	0.039*** (0.004)
No EPP	−0.034*** (0.013)	−0.047*** (0.012)	−0.282*** (0.040)	−0.023*** (0.008)	0.001 (0.005)
Customer Service	−0.160*** (0.038)	−0.016 (0.032)	−0.484*** (0.119)	−0.108*** (0.024)	−0.036** (0.016)
Log(Num. Prior Bookings)	0.335*** (0.003)	0.034*** (0.002)	0.869*** (0.008)	0.542*** (0.002)	0.038*** (0.001)
Log(First Price)	−0.140*** (0.003)	0.339*** (0.003)	−0.298*** (0.010)	−0.140*** (0.002)	0.014*** (0.001)
<i>Treat</i> × >Median EPP	−0.005 (0.013)	0.0005 (0.010)	−0.005 (0.042)	−0.008 (0.008)	−0.005 (0.006)
<i>Treat</i> × No EPP	0.004 (0.016)	0.012 (0.015)	−0.007 (0.051)	0.002 (0.010)	0.002 (0.007)
<i>Treat</i> × Customer Service	0.034 (0.054)	0.039 (0.046)	0.076 (0.170)	−0.043 (0.034)	−0.031 (0.022)
Observations	119,550	73,234	119,550	119,550	119,550
R ²	0.262	0.411	0.219	0.631	0.078

Notes. This table displays the treatment effects on listing outcomes after the first transaction in the experiment. Controls are included for greater-than-median EPP (>Median EPP), whether the EPP is calculable, log of prior bookings, log of the first price, and whether the guest submitted a customer service complaint. Columns with “in Exp.” in the dependent variable refer to outcomes calculated only through June 12, 2014, the end of the experimental period. In column 3, “Rev.” refers to the revenue of the listing during the experimental period. There are fewer observations for the price variable. This is because we cannot measure transaction prices for hosts who did not transact after the initial transaction in the experiment.

** $p < 0.05$; *** $p < 0.01$.

meaning that it is a good proxy of listing quality. However, the interaction of this variable with the treatment is close to zero and not statistically significant. We conduct a similar exercise with another proxy of quality—the occurrence of a customer service complaint during the first transaction in the experiment.¹⁶ We find that customer service complaints are associated with worse subsequent listing outcomes even in the control. However, we find no statistically significant interaction effects with the treatment. To summarize, we do not find decreases in adverse selection using two proxies for listing quality.

One may also be concerned that the small average treatment effects mask other types of heterogeneity not identified by our proxies for low-quality listings. For example, a marginal positive or negative review may have large effects for a subset of listings. We test for this by comparing the distribution of bookings through 2014 between the treatment and control. Figure A4 in the online appendix shows that these distributions are very similar. We also conduct a Kolmogorov–Smirnov test on the equality of these distributions and fail to reject the null ($p = 0.6296$). This confirms that earlier exposure to the treatment had, at most, negligible effects on average market outcomes. We discuss the implications of these findings in the next section.

10. Discussion

Reputation systems are an important component of a well-functioning online marketplace. However, because informative reviews are public goods, reputation systems do not capture all relevant information, and observed ratings may be biased. These systems may be especially difficult to design for peer-to-peer markets in which services are exchanged. In these settings, market participants can review each other and may meet in person, resulting in reciprocity and retaliation within the review system. To our knowledge, all major platforms with two-sided review systems have implemented systems where users are unable to see their counterparty’s review before writing their own. However, reviews are unveiled to the reviewer only on some platforms (Upwork and Freelancer) but not on others (Lyft and Uber). In this paper, we study the effects of a simultaneous reveal policy intended to reduce reciprocity and to improve market outcomes. Our results suggest that whether the review is unveiled plays a critical role in the effects of the simultaneous reveal design.

We find that the simultaneous reveal policy increased review rates and decreased the average valence of reviews. It also reduced retaliatory one-star reviews as well as the correlation between guest and

host ratings. The effects we find are due to at least two factors: a reduction in reciprocity and what we refer to as the desire to unveil reviews. We also note that although the relative effects of the treatment on reviews are substantial—the treatment increased reviews with negative text by over 12% for both guests and hosts—the absolute effects are small. For example, negative review text by guests occurs in just 8.7% of transactions in the control, so only a small share of transactions are affected by this treatment.

The ultimate goal of reputation system changes should be to improve the quality of transactions in the market. For example, the intention of the simultaneous reveal policy was to make reviews more commensurate with experienced transaction quality, with the idea that more informative reviews will lead to better matches. Of the factors we document, the reduction in reciprocity should indeed have this intended effect. On the other hand, the informative value of additional reviews induced by the desire to reveal review information is uncertain. We study whether simultaneous reveal led to better matches and reduced adverse selection and find that it did not. Note that the null effects we find may be driven by the fact that the experiment ran for a relatively short period of time prior to simultaneous reveal reviews being launched across the entire site.

We draw several other lessons about reputation systems from our results. First, although it is widely known that review information can be biased, it is less acknowledged that magnitude of this bias can change over time due to changes in the reputation system design. This can be true even for aspects of the review that are anonymous and/or private and, consequently, expected to be less subject to bias. The simultaneous reveal treatment only affected the timing of the disclosure of review text to a counterparty. Nonetheless, the treatment changed both review text and star ratings.

Another lesson we draw is that real world reviewing behavior may be hard to replicate in a laboratory setting. The laboratory tests of the simultaneous reveal policy conducted by Bolton et al. (2012) showed decreases in review rates whereas we found increases. We show that this can be explained by the desire to unveil reviews, a motivation for reviewing not present in the laboratory experiment. Other potentially important differences between our setting and the laboratory include differences in the underlying distribution of transaction quality and the presence of social, rather than strategic, reasons for submitting high ratings.

We do not exhaustively study the determinants of Airbnb's ratings distribution. For instance, social interactions before, during, or after a stay on Airbnb may lead market participants to omit relevant

information from their reviews. Furthermore, not all users submit reviews on Airbnb. If those that opt out of reviewing have lower quality experiences, reviews on the platform will tend to be more positive. Our principal stratification results demonstrate that who selects into reviewing can affect the observed rating distribution. It is also possible that reviewers leave different types of feedback when they know their name and account will be publicly associated with review text. There is room to explore designs that allow reviewers to opt out of associating their reviews with their profiles.

The ratings distribution is also influenced by platform enforcement actions including listing removals and penalties in search rankings. For example, Airbnb's trust and safety team has filtered approximately 970,000 problematic listings from the platform (Swisher 2019). We do not know the importance of these actions.

Finally, reviews may describe how an experience compared with the reviewer's own expectations, rather than describing an experience's absolute quality. For example, for cheaper Airbnb listings, guests may not expect hotel-quality amenities and service from the host. It should be possible to design review systems that separate expectation-based ratings from more objective evaluations. Indeed, Airbnb has tried to create this separation by asking guests about specific features of a home and grouping listings by those features. "Airbnb Plus" homes not only have high ratings, but are also visited in person by an Airbnb representative to ensure quality, amenities, and the accuracy of the listing description. Similarly, "For Work" homes are those that have WiFi, a work space, and self check-in. The extent to which these complementary reputation mechanisms affect market outcomes remains a question for future work.

Acknowledgments

The authors thank the editor, Avi Goldfarb, and the referees for providing suggestions that improved this paper. They also thank Chris Dellarocas, Dean Eckles, Liran Einav, Chiara Farronato, Shane Greenstein, John Horton, Caroline Hoxby, Ramesh Johari, Max Kasy, Jon Levin, Tetsy Lin, Mike Luca, Jeff Naecker, Fred Panier, Catherine Tucker, Giorgos Zervas, and seminar participants at Microsoft, the Massachusetts Institute of Technology, eBay, the University of Hong Kong, the 16th ACM Conference on Economics and Computation, the National Bureau of Economic Research Summer Institute, Conference on Digital Experimentation @ MIT, and the Advances in Field Experiments Conference for comments. They thank Matthew Pearson for early conversations regarding this project and Peter Coles and Mike Egedal 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.

Endnotes

¹ Reciprocity has also been studied in other digital settings. Lorenz et al. (2011) use an experiment to show how adding social information to a wisdom of crowds task increases bias, and Livan et al. (2017) find evidence of reciprocity in content platforms.

² There is also an incentive in the RIT system to wait until the other party reviews in order to threaten retaliation. Our data does not suggest that this is an important motivation on Airbnb. Hosts typically review first and more positively than guests. This is true even though hosts have much more to lose from a negative review than guests (Cui et al. 2020). We interpret this as evidence that most hosts value the benefit of inducing a positive review more than the benefit of waiting to threaten a negative review.

³ There were some cases where a review was submitted after the 14- or 30-day time period. This occurred because of the manner in which emails were batched relative to the time zone, modifications to the trip parameters, or bugs in the review prompt system.

⁴ In the mobile app, the stars are labeled (in ascending order) “terrible,” “not great,” “average,” “great,” and “fantastic.” The stars are not labeled on the main website during most of the sample period.

⁵ See Online Appendix A for additional details on the logging of review-related data.

⁶ A final third were assigned to the status quo before the experiment, in which reviews were released as soon as they were submitted and there was a 30-day review period. We do not focus on the status quo in this paper because the difference in the reviewing period may have had an effect separate from the simultaneous revelation of reviews.

⁷ We do this because we are, for the time being, interested in the effects of the experiment on reviewing behavior rather than on adverse selection, which may affect subsequent transactions and reviews. Because guests and hosts in this sample did not know about the change to the review system before the trip, they cannot adjust their match to the new policy. In contrast, for subsequent transactions, the treatment may affect selection into transactions. Furthermore, this sample restriction allows us to avoid issues due to spillovers between multiple listings managed by the same host.

⁸ Although randomization began for trips ending on May 7, 2014, we exclude trips with checkouts between May 7 and May 9, 2014, because of inconsistencies in logging treatment assignments on those days. Online Appendix A recreates our main results with a sample that excludes any host with a trip ending on these days. This appendix also includes details regarding treatment assignment logging issues on June 6 and June 7, 2014. Because we analyze only each host’s first trip during the experiment and this span of days occurs toward the end of the experiment, these logging issues do not substantively affect our results. Note that the experiment ran all the way until the public announcement and launch of the policy to the entire platform. We do not use data from close to the launch in our main analysis because reviewing behavior may have been affected by the launch.

⁹ We find similar results in a linear model where the outcome variable is whether a review by a user comes within a day after a review by the counterparty (Table AIII in the online appendix).

¹⁰ The fact that even in the control group a first review speeds up the second may be explained by one of three factors. First, the first review may serve as a reminder. Second, the first review may induce a reciprocal obligation to review. Last, the speed of guest and host reviews may be correlated with each other due to unobserved heterogeneity.

¹¹ An analogous assumption regarding never reviewers and compliers in the control holds trivially because they do not submit reviews.

¹² The composition of ratings submitted by compliers is displayed in Figure A2 in the online appendix. The ratings left by compliers are typically lower than those of always reviewers.

¹³ We can also reject large effects of the treatment on subsequent guest outcomes (Table AIV in the online appendix).

¹⁴ Klein et al. (2016) propose a toy model of reviewing, retaliation, and market outcomes. In their model, eliminating retaliation induces more honest (lower) ratings and causes seller exit and increased effort provision. Because our treatment had effects on review quantity and speed in addition to reducing average ratings, these simple predictions do not necessarily apply to our setting.

¹⁵ Another possibility is that the treatment reduces moral hazard, which we were unable to test for because we cannot measure quality. Using realized ratings, in the treatment, as measures of quality, as in the prior literature, is problematic because the treatment affects ratings in ways other than through quality.

¹⁶ We exclude customer service complaints that occur after the transaction has finished because they may be affected by the treatment.

References

- Airbnb (2014) Building trust with a new review system—The Airbnb Blog—Belong Anywhere. Accessed August 4, 2021, <https://blog.airbnb.com/building-trust-new-review-system/>.
- Avery C, Resnick P, Zeckhauser R (1999) The market for evaluations. *Amer. Econom. Rev.* 89(3):564–584.
- Bolton G, Greiner B, Ockenfels A (2012) Engineering trust: Reciprocity in the production of reputation information. *Management Sci.* 59(2):265–285.
- Bondi T (2019) Alone, together: Product discovery through consumer ratings.
- Brandes L, Godes D, Mayzlin D (2020) What drives extremity bias in online reviews? Theory and experimental evidence. Working paper, University of Lucerne, Lucerne, Switzerland.
- Cabral L, Hortaçsu A (2010) The dynamics of seller reputation: Evidence from Ebay. *J. Indust. Econom.* 58(1):54–78.
- Cabral LMB, Li L (2012) A dollar for your thoughts: Feedback-conditional rebates on Ebay. *Management Sci.* 61(9):2052–2063.
- Cui R, Li J, Zhang DJ (2020) Reducing discrimination with reviews in the sharing economy: Evidence from field experiments on Airbnb. *Management Sci.* 66(3):1071–1094.
- Dellarocas C, Wood CA (2007) The sound of silence in online feedback: Estimating trading risks in the presence of reporting bias. *Management Sci.* 54(3):460–476.
- Ding P, Lu J (2017) Principal stratification analysis using principal scores. *J. Roy. Statist. Soc. Ser. B. Statist. Methodol.* 79(3):757–777.
- Dunbar RIM, Marriott A, Duncan NDC (1997) Human conversational behavior. *Human Nature* 8(3):231–246.
- Filippas A, Horton JJ, Golden J (2018) Reputation inflation. *Proc. 2018 ACM Conf. Econom. Comput.* (Association for Computing Machinery, New York), 483–484.
- Fradkin A, Grewal E, Holtz D, Pearson M (2015) Bias and reciprocity in online reviews: Evidence from field experiments on Airbnb. *Proc. 16th ACM Conf. Econom. Comput.* (Association for Computing Machinery, New York), 641–641.
- Frangakis CE, Rubin DB (2002) Principal stratification in causal inference. *Biometrics* 58(1):21–29.
- Hui X, Saeedi M, Sundaresan N (2018) Adverse selection or moral hazard: An empirical study. *J. Indust. Econom.* 66(3):610–649.
- Jaffe S, Coles P, Levitt S, Popov I (2019) Quality externalities on platforms: The case of Airbnb. Working paper, Microsoft’s Office of the Chief Economist.
- Klein TJ, Lambert C, Stahl KO (2016) Market transparency, adverse selection, and moral hazard. *J. Political Econom.* 124(6):1677–1713.

- Lafky J (2014) Why do people rate? Theory and evidence on online ratings. *Games Econom. Behav.* 87:554–570.
- Levitt SD, List JA (2007) What do laboratory experiments measuring social preferences reveal about the real world? *J. Econom. Perspect.* 21(2):153–174.
- Li X, Hitt LM (2008) Self-selection and information role of online product reviews. *Inform. Systems Res.* 19(4):456–474.
- Livan G, Caccioli F, Aste T (2017) Excess reciprocity distorts reputation in online social networks. *Sci. Rep.* 7(1):3551.
- Loewenstein G (1994) The psychology of curiosity: A review and interpretation. *Psych. Bull.* 116(1):75–98.
- Lorenz J, Rauhut H, Schweitzer F, Helbing D (2011) How social influence can undermine the wisdom of crowd effect. *Proc. Natl. Acad. Sci. USA* 108(22):9020–9025.
- Miller N, Resnick P, Zeckhauser R (2005) Eliciting informative feedback: The peer-prediction method. *Management Sci.* 51(9):1359–1373.
- Nosko C, Tadelis S (2015) The limits of reputation in platform markets: An empirical analysis and field experiment. Working paper, Uber.
- Proserpio D, Xu W, Zervas G (2018) You get what you give: Theory and evidence of reciprocity in the sharing economy. *Quant. Marketing Econom.* 16(4):371–407.
- Silvia PJ (2012) Curiosity and motivation. Ryan RM, ed. *The Oxford Handbook of Human Motivation* (Oxford University Press, New York), 157–166.
- Swisher K (2019) Brian Chesky: How Airbnb is responding to a deadly shooting. *Recode Decode* (podcast). (November 8, 2019), <https://podcasts.apple.com/us/podcast/decoder-with-nilay-patel/id1011668648>.
- Zervas G, Proserpio D, Byers J (2021) A first look at online reputation on Airbnb, where every stay is above average. *Marketing Lett.* 32:1–16.