QUALITY CERTIFICATIONS INFLUENCE USER-GENERATED RATINGS

ABSTRACT

Platforms present various certifications to signal the quality of their offerings to prospective consumers. For example, Airbnb.com designates some hosts as “Superhosts” to distinguish properties that provide superior experiences. Platforms also present user-generated ratings—typically elicited and presented as “star ratings”—from their customers for the same purpose. This research investigates the interaction of these signals of quality and suggests a potential downside to platform-provided certifications: They decrease subsequent ratings. In an analysis of over 1,500,000 ratings from Airbnb.com and three follow-up studies, we find that properties with the superhost designation receive lower ratings. We assess the robustness of this result in several ways, including comparing ratings on Airbnb with those for the same property of Vrbo. In three follow-up experiments, we find that the net effect of certifications can lead to reduced choice share: The positive effect of signaling quality is more than offset by the negative effect of reduced ratings. This suggests that consumers are not sufficiently aware of this effect of quality certifications on ratings when choosing.

Keywords: *user-generated ratings*, *online ratings*, *evaluability*

INTRODUCTION

E-commerce platforms offer a variety of information about products to help consumers make informed choices when shopping online. For example, technical product specifications are nearly universally displayed online. Likewise, many platforms present platform-specific certifications. Airbnb.com signals high quality listings through “Superhost” status, eBay.com awards sellers with a “Top Rated Seller” designation, and Apple promotes certain apps as “Editor’s Picks” and “Apps We Love.” Platforms create these certifications as signals of quality, hoping to stimulate demand. Past research suggests that these signals work as intended. For example, Airbnb’s superhost status has been shown to increase bookings for designated listings (Yao et al. 2019) and, more generally, to increase overall bookings on the platform (Mishra, Huang and Kalwani 2023). Similarly, eBay’s top rated seller designation has been shown to increase demand for designated sellers (Elfenbein, Fisman and McManus 2015; Hui et al. 2016; Lewis 2011; Li, Srinivasan and Sun 2009).

Meanwhile, platforms also provide user-generated ratings (e.g., star ratings on Amazon.com) for consumers. User-generated ratings are a ubiquitous feature of the online consumer experience, and research suggests that consumers trust them. Consumers are reluctant to buy products without ratings (Askalidis, Kim and Malthouse 2017), and when comparing multiple options, consumers tend to purchase products with higher ratings (Chen, Wang and Xie 2011; Chintagunta, Gopinath and Venkataraman 2010; Dellarocas, Zhang and Awad 2007), likely because they expect higher-rated options to give them more utility (de Langhe, Fernbach and Lichtenstein 2016).

While both platform-created certifications and user-generated ratings increase demand for awarded or highly-rated alternatives, neither exists in isolation. Instead, it is likely that platforms' certifications affect users' ratings. This is because certifications provide a context within which consumers create their ratings. To help illustrate this idea, consider the ratings for Pulitzer Prize winning books on Goodreads.com. The average prize winner receives a rating barely above average ($M_{\text{PrizeWinners}}$ = 4.00/5 vs. M_{AllBooks} = 3.89/5) and ranks in the 59th percentile of all books in its publication year (see Web Appendix A for more details). Presumably, this is not because Pulitzer Prize winning books are of middling quality, but because prize winning books are rated within the context of being "the best book of the year." Non prize winners are unlikely to be rated against such a high bar.

This raises several questions about the impact of platform-created certifications on ratings. First, do platforms' certifications lead to diminished ratings? The Goodreads example provides anecdotal evidence that this may be the case. A second question is contingent on the first: If quality-signaling certifications dampen ratings, what impact does this have on consumers' choices? If prospective consumers are aware of the impact of quality signals on ratings, there is little reason for concern. However, if prospective consumers are not aware of this impact, or are aware but insufficiently adjust for it, it should warrant some concern. In this scenario, the effectiveness of platform-created signals in stimulating demand would be diminished, as part of their positive impact would be offset by the consequent reduction in user-generated ratings.

This manuscript explores these questions in real-world data and follow-up lab experiments. Our results suggest that platform-created certifications (e.g., quality designations) can reduce user-generated ratings for certified products. Further, we find that prospective consumers are aware of the effect of certifications on ratings to some extent, but insufficiently so. Instead, prospective consumers are apt to mischaracterize differences in ratings as reflecting differences in quality, even when those ratings are affected by certifications.[1] As a result, our findings suggest that platforms' signals of quality are likely less effective than intended. While they stimulate demand in isolation, their dampening effect on ratings has the opposite effect. This is because many consumers assume that the higher-rated alternative must be better. This is true even when they have the information necessary to understand the true cause of the rating difference.

We reach these conclusions through a multi-method, "data rich" investigation (Blanchard et al. 2022). First, in Study 1, we use field data of over 1,500,000 ratings from the peer-to-peer homesharing platform Airbnb.com to assess the effect of a platform-created certification on ratings. Results suggest that when a property is honored with the distinction of superhost, consumers rate the property more harshly, giving it lower ratings than if it had not been given the distinction. We contend this is because the superhost designation provides context, and the ratings consumers create depend on that context, such that superhosts are compared to increased expectations and/or higher-quality alternatives. Next, we present the results of three laboratory studies. First, we replicate the Airbnb result in an experimental context, where we can exogenously vary the presence/absence of the superhost certification. Then, in Studies 2A and 3, we examine the joint effect of the superhost certification and the diminished ratings it entails in a choice context.

[1] We thank an anonymous reviewer for the wording of this synopsis.

CONCEPTUAL DEVELOPMENT

Because consumers cannot see, touch, feel, or experience most of their offerings, online platforms have to curate information for consumers. This includes platform-specific certifications and designations, designed to simplify information and clearly signal high-quality offerings. Examples of certifications are superhost status on Airbnb.com, which Airbnb uses to identify hosts who go "above and beyond to provide excellent hospitality" (Airbnb 2024), eBay's top rated seller designation, which indicates sellers who similarly excel, Apple's "Apps We Love," Indigo Bookstores' "Heather's Picks," and Kickstarter's "Projects We Love." Prior research suggests these certifications can increase demand (Elfenbein et al. 2015; Fleischer, Ert, and Bar-Nahum 2022; Hui, Lui, and Zhang 2023; Hui et al. 2016; Lewis 2011; Li et al. 2009; Mishra et al. 2023; Yao et al. 2019).[2] However, prior research has not considered the effect of certifications on consumers' evaluations of their experiences—often expressed through user-generated ratings.

As the primary form of information generated by other consumers, ratings are uniquely capable of communicating in simple terms what something is "like to own" (Simonson 2016). Ideally, these ratings are an unbiased source of information upon which prospective consumers can compare alternatives. If a consumer has a good experience with a product, they should rate it five-stars. Bad experience? They should provide a lower rating. Thus, averaging many past consumers' ratings for a product or service should provide an unbiased representation of the average experience one can expect. This notion is in fact consistent with the way consumers use

[2] Outside of purchase contexts, Rietveld, Seamans, and Meggiorin (2021) found that microfinance lenders saw an increase in demand for loans after receiving a "social certification" badge.

ratings—to compare competing alternatives with the aim of deciding which option in a set is best to purchase (de Langhe et al. 2016). Unfortunately, it is unlikely that ratings live up to this ideal of being an unbiased point of comparison across alternatives. Specifically, we contend that certifications influence the context in which consumers make ratings, leading to more negative ratings for certified products and services.

Impacts of Platform Certifications on Ratings

Consumers' ratings for products follow a similar cognitive process to any judgment they make. In general, people form judgments by using information that is explicitly presented (Slovic 1972), or readily available when information is not presented (Lynch, Marmorstein and Weigold 1988). When creating ratings online, very little information is explicitly present, requiring consumers to make ratings by bringing their own information to mind. We contend that platform-created certifications bring different information to mind for certified and uncertified experiences. Specifically, offerings signaled as high quality by platforms will be compared to increased expectations, or exceptional remembered or idealized experiences; for example, consumers asked to rate a superhost Airbnb property might compare their experience to other superhosts they have stayed at, or what they imagine a superhost to be, while consumers asked to rate a non-superhost might compare their experience to other non-superhosts.

The expectation-disconfirmation and service quality literatures show that consumers' evaluations are a result of the alignment of a consumer's experience with their expectations (Bearden and Teel 1983; Oliver 1977, 1980; Parasuraman, Zeithaml and Berry 1985; 1988), with these expectations often arising from attributes of the experience itself (e.g., brands, marketing material, prior experience; Woodruff, Cadotte and Jenkins 1983). All else equal, higher expectations lead to lower ratings. This is consistent with a nascent literature in quantitative marketing on *critic-awarded* certifications, which argues for a negative effect of Michelin stars (Li et al. 2022) and Academy Award nominations (Bondi and Stevens 2019; Rossi 2021) on ratings.

This body of research has clear implications for the possible effect of platform-created certifications on ratings. These signals may lead consumers to compare their experience against higher expectations or higher-quality alternatives. Meanwhile, these signals have no impact on actual quality—an Airbnb does not automatically become higher quality when it receives the superhost designation. Thus, we propose:

H_1: User-generated ratings will be lower when platform certifications signal a product or service to be of high quality.

Consumers' Interpretation of User-Generated Ratings

There is nothing inherently wrong with users creating ratings differently for products with platform certifications of quality (H_1): If prospective consumers realize that superhosts are judged on a harsher scale than non-superhosts, they can interpret the observed ratings appropriately. However, an issue arises if prospective consumers—when they are interpreting these ratings—fail to recognize, or insufficiently adjust for, the influence of platform certifications of quality on ratings.

Whether prospective consumers are aware of what influenced the ratings they observe remains an open question. For example, the aforementioned literatures on expectation-disconfirmation and service quality do not analyze this question. It has been argued that consumers are aware of the influence of expectations on their own ratings (Churchill Jr and Surprenant 1982; Grönroos 1982; Lewis and Booms 1983; Parasuraman et al. 1985). However, to our knowledge, this literature has not found that people recognize the influence of expectations on other consumers' evaluations.

Meanwhile, consumer research outside of expectation-disconfirmation and service quality suggests that consumers—when comparing the ratings of multiple products—are unlikely to spontaneously consider the information those prior raters used. For one, decision makers often take the information they are given at face value (i.e., what you see is all there is; Kahneman 2011), making them unlikely to consider the hidden information that led to a rating. Instead, user-generated ratings have many properties that contribute to their unscrutinized use. They are readily available, being presented explicitly in the environment, and easy to evaluate due to their ubiquity (Kivetz and Simonson 2000; Lynch et al. 1988; Nowlis and Simonson 1997; Slovic 1976; Slovic and MacPhillamy 1974). Thus:

H_2: Prospective consumers do not sufficiently correct for the influence of platform certifications when interpreting ratings.

We believe the contribution of this manuscript is in the combination of hypotheses. Together, these two hypotheses suggest that platforms' signals of quality are inefficient drivers of demand. While past research has shown that signals increase demand in isolation, this effect is diminished if they decrease ratings, and if consumers over-rely on ratings.

EMPIRICAL INVESTIGATION

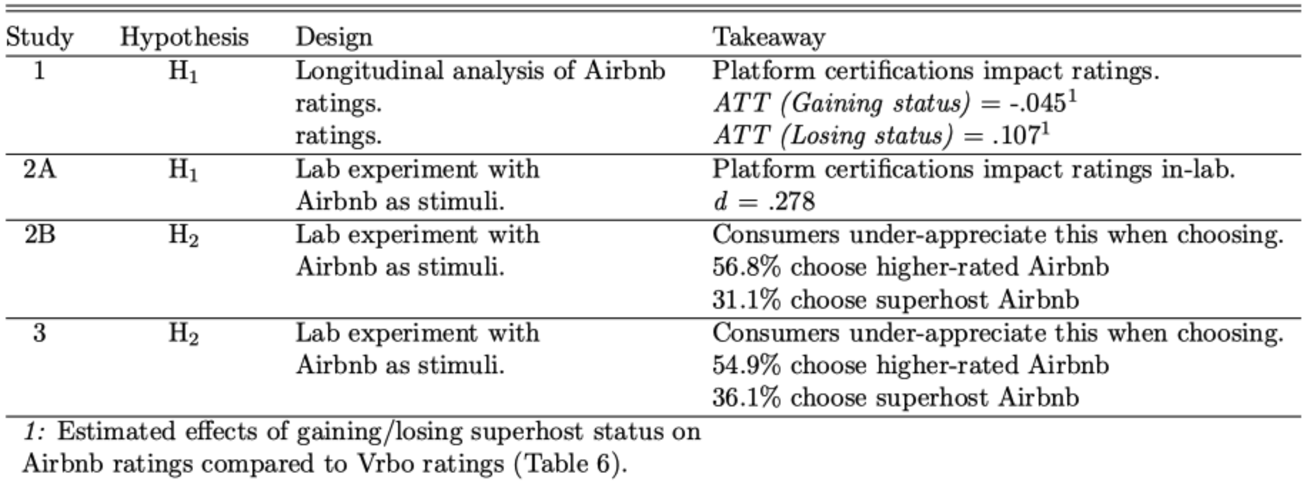
The following empirical investigation includes three types of data: (i) rating data from real e-commerce platforms (from Airbnb and Vrbo; Study 1), (ii) rating data from a laboratory experiment (Studies 2A), and (iii) choice data from laboratory studies (Studies 2B and 3). These different types of data complement each other. For example, while the real rating data offers ecological validity, it presents challenges in terms of unambiguous identification of causal effects. The lab data, on the other hand, allows straightforward causal inference, but lacks the richness associated with consumption behavior in the wild. Additionally, the combination of studies assessing ratings and those assessing choice allow us to examine the interplay of H_1 and H_2: If quality-signaling certifications lower consumer ratings, do prospective consumers realize this when they are interpreting these ratings to make choices?

To preview the results, first, in Study 1, we examine whether platform certifications affect user-generated ratings in real markets, analyzing rating data from Airbnb. We find that possessing the superhost certification is associated with lower ratings: When a property gains the superhost designation its ratings subsequently get worse and when a property loses the superhost designation its ratings subsequently get better. Our results suggest these changes in ratings are not caused by changes in quality, as the ratings for the same properties on an alternative platform (Vrbo) are not affected. Instead, we argue that ratings drop when a property receives superhost status because expectations go up and the ratings are provided conditional on these newly inflated expectations.

Next, in Study 2A we examine the effect of the superhost certification on ratings in a controlled laboratory experiment. We replicate the results of Study 1: The same property receives lower ratings when it has (versus does not have) the superhost certification. Finally in Studies 2B and 3, we examine how prospective consumers interpret ratings that have been affected by platform certifications. We find consumers are insufficiently attentive to the decrease in ratings caused by quality-signaling certifications and instead chose as if star ratings were an unbiased measure of quality.

Table 1 summarizes the design and conclusions from all studies in this manuscript. All laboratory studies were pre-registered. All code, data, materials, and pre-registrations (including code used to collect Vrbo and Airbnb ratings) are available on our OSF repository (https://osf.io/3he6c/?view_only=e031a89ca6fd464ebb67de90e0363014).

TABLE 1
SUMMARY OF STUDIES



Study	Hypothesis	Design	Takeaway
1	H_1	Longitudinal analysis of Airbnb ratings. ratings.	Platform certifications impact ratings. *ATT (Gaining status)* = -.045[1] *ATT (Losing status)* = .107[1]
2A	H_1	Lab experiment with Airbnb as stimuli.	Platform certifications impact ratings in-lab. d = .278
2B	H_2	Lab experiment with Airbnb as stimuli.	Consumers under-appreciate this when choosing. 56.8% choose higher-rated Airbnb 31.1% choose superhost Airbnb
3	H_2	Lab experiment with Airbnb as stimuli.	Consumers under-appreciate this when choosing. 54.9% choose higher-rated Airbnb 36.1% choose superhost Airbnb

1: Estimated effects of gaining/losing superhost status on Airbnb ratings compared to Vrbo ratings (Table 6).

STUDY 1: THE EFFECT OF SUPERHOST STATUS ON AIRBNB RATINGS

In Study 1 we assess H_1: That platform-hosted, quality-signally certifications will lead to lower ratings for certified alternatives. To do this, we investigate Airbnb.com—an online marketplace for peer-to-peer home rentals. When consumers browse listings on Airbnb, one of the many pieces of information they see is the superhost designation, which Airbnb claims to use to identify hosts who go “above and beyond to provide excellent hospitality” (Airbnb 2024). Airbnb claims to award superhost status to hosts who have (i) earned an average rating of 4.8/5 or above, (ii) responded to at least 90% of guests within 24 hours, (iii) hosted at least 10 stays, and (iv) canceled 1% of bookings or less, all in the last year. These criteria are evaluated on January 1, April 1, July 1, and October 1 each year. If hosts become superhosts, they are awarded with a small badge displayed to consumers on their listings’ pages. Thus, the superhost designation is a host-level designation, which appears on individual property pages.

We predict that, all else equal, this superhost designation lowers user-generated ratings. However, we cannot simply compare average ratings between properties from superhosts to those from non-superhosts. This is because superhost status is awarded on merit, not randomly assigned. Thus, higher quality listings are more likely to be superhosts than low quality listings. So, a between-listing comparison between the superhost and non-superhost ratings would not just reflect the effect of the superhost label, but would also reflect the differences in quality that lead to possession of the superhost label.

We present three alternative identification strategies in attempt to mitigate this lack of random assignment. The first is a difference-in-differences design, comparing how ratings change over time for listings who obtain or lose superhost status versus those whose superhost status does not change. While this straightforward analysis supports our prediction, caution is warranted when interpreting the results, as listing owners have direct control over their treatment (superhost) status. The second identification strategy utilizes fixed-effect regression and avoids between-group selection issues by focusing only on within-listing differences in superhost status.

This analysis also allows us to control for selection of raters, as we can remove between-rater variation with rater fixed effects. The third identification strategy compares ratings for the same property across platforms: We find that superhost status on Airbnb does not systematically affect ratings on Vrbo, consistent with our expectations-based hypothesis but inconsistent with alternative accounts that rely on time-varying quality.

Results of all identifications support H_1: We observe that superhost status leads to lower ratings. The following section introduces our data, including a discussion of the frequency and probability of changing superhost status. Then, we present model-free evidence for the effect of superhost status on ratings. We then introduce our three identification strategies; (i) difference-in-differences in Airbnb ratings across listings, (ii) within-listing analysis of ratings, and (iii) difference-in-differences in Airbnb and Vrbo ratings. After introducing each, we present results including robustness analyses.

Data

Our Airbnb data come from three sources; (i) quarterly snapshots of Airbnb listings for six quarters between September 2021 and December 2022 collected from InsideAirbnb.com, (ii) individual ratings and reviews for those listings between July 2021 and December 2022 that we collected from Airbnb, and (iii) individual ratings from Vrbo listings during the same time period.

InsideAirbnb Data. We obtained a panel of 1,420,922 quarterly observations of 405,765 American Airbnb listings from InsideAirbnb.com for the six quarters between September 2021 and December 2022. Each observation is a snapshot of a listing's customer-facing page at the time of collection. Most importantly for our purposes, each quarterly observation includes whether or not the listing had superhost status for that quarter, as well as hosts' response rate in the 30 days before each snapshot, listings' amenities, number of reviews, price, and number of people accommodated. We focus on the 133,706 listings that have ratings across more than one quarter. By examining these listings over time, we identify listings that are always superhosts (40,311; 30.1% of total), never superhosts (44,460; 33.3%), or have variation in superhost status (48,937; 36.6%). Those with variation in superhost status can further be segmented into listings who gain status and never lose it (24,461; 50.0% of those with variation), lose status and never regain it (11,632; 23.8%), or both gain and lose it (12,846; 26.3%).[3]

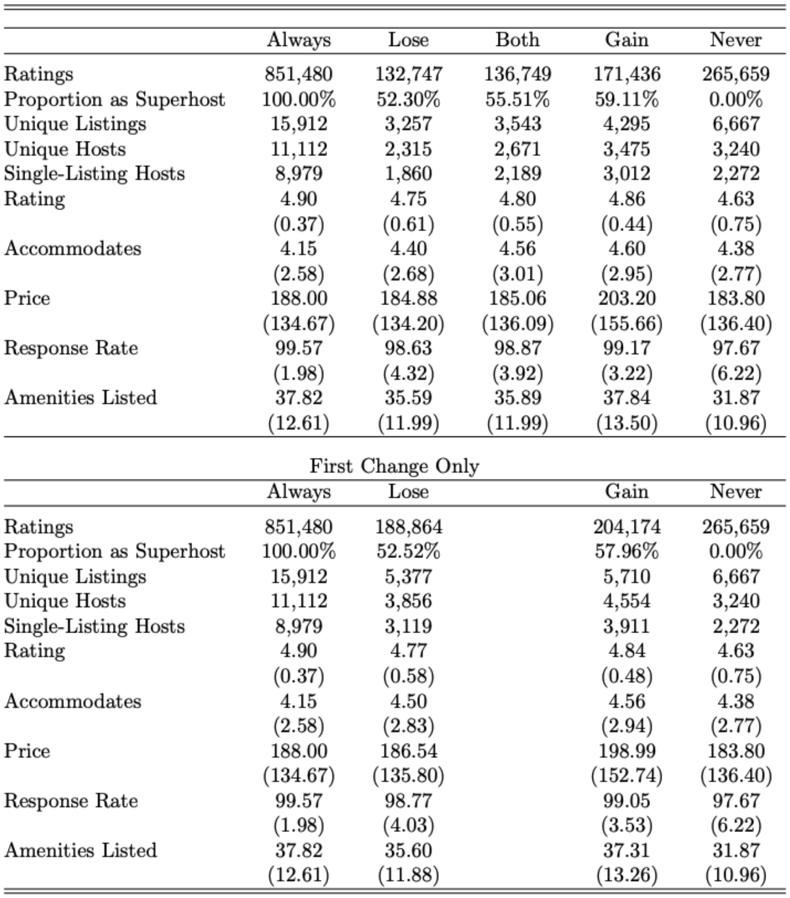
Individual Airbnb Ratings. While the InsideAirbnb data includes each listing's average rating at the time of observation, it does not include individual ratings. To supplement the InsideAirbnb data, we obtained these individual ratings—including the rating level (1–5), date, and reviewer ID—directly from Airbnb in June 2024. This resulted in 1,558,071 individual ratings from 33,674 unique listings and 1,389,461 unique raters.

The resulting set of listings for which we have individual ratings is smaller and of slightly different composition than the full InsideAirbnb set. We have a higher proportion of listings who are always superhosts (47.3% vs 30.1% in the initial sample), and slightly fewer who are never superhosts (19.8% vs 33.3%) or have variation in status (32.9% vs 36.6%). Within the subset of

[3] Snapshots were collected immediately before superhost status changes (Web Appendix B). Thus, whatever information changes between each listing snapshot likely changed under the superhost status observed in that snapshot. This also means that the superhost status in the most recent snapshot is the same as what each rater saw when evaluating the property.

listings with variation in superhost status, we have more listings who gain (38.7%, vs 50.0% in the initial sample), and relatively similar proportions of those who lose (29.4%, vs 23.8%), or do both (31.9%, vs 26.3%). This discrepancy is because we could not collect individual ratings for listings who left Airbnb between our final InsideAirbnb observation and summer 2024.[4]

TABLE 2
DESCRIPTIVES OF AIRBNB LISTINGS UNDER DIFFERENT SUPERHOST STATUSES



	Always	Lose	Both	Gain	Never
Ratings	851,480	132,747	136,749	171,436	265,659
Proportion as Superhost	100.00%	52.30%	55.51%	59.11%	0.00%
Unique Listings	15,912	3,257	3,543	4,295	6,667
Unique Hosts	11,112	2,315	2,671	3,475	3,240
Single-Listing Hosts	8,979	1,860	2,189	3,012	2,272
Rating	4.90	4.75	4.80	4.86	4.63
	(0.37)	(0.61)	(0.55)	(0.44)	(0.75)
Accommodates	4.15	4.40	4.56	4.60	4.38
	(2.58)	(2.68)	(3.01)	(2.95)	(2.77)
Price	188.00	184.88	185.06	203.20	183.80
	(134.67)	(134.20)	(136.09)	(155.66)	(136.40)
Response Rate	99.57	98.63	98.87	99.17	97.67
	(1.98)	(4.32)	(3.92)	(3.22)	(6.22)
Amenities Listed	37.82	35.59	35.89	37.84	31.87
	(12.61)	(11.99)	(11.99)	(13.50)	(10.96)

First Change Only

	Always	Lose		Gain	Never
Ratings	851,480	188,864		204,174	265,659
Proportion as Superhost	100.00%	52.52%		57.96%	0.00%
Unique Listings	15,912	5,377		5,710	6,667
Unique Hosts	11,112	3,856		4,554	3,240
Single-Listing Hosts	8,979	3,119		3,911	2,272
Rating	4.90	4.77		4.84	4.63
	(0.37)	(0.58)		(0.48)	(0.75)
Accommodates	4.15	4.50		4.56	4.38
	(2.58)	(2.83)		(2.94)	(2.77)
Price	188.00	186.54		198.99	183.80
	(134.67)	(135.80)		(152.74)	(136.40)
Response Rate	99.57	98.77		99.05	97.67
	(1.98)	(4.03)		(3.53)	(6.22)
Amenities Listed	37.82	35.60		37.31	31.87
	(12.61)	(11.88)		(13.26)	(10.96)

Table 2 provides descriptive statistics of these combined data sources. Apart from average ratings and the number of reviews, the properties with variation in superhost status look largely similar to the other two groups. In our difference-in-difference models (i.e., the first and

[4] This should actually improve the internal validity of these data, as listings who left the platform are more likely to have changes in quality over time.

third identification strategies), we consider only the first change in status for the listings who both gain and lose status. We split each listing into the relevant group, and then drop observations from after the second change in status. The bottom panel of Table 2 presents descriptive statistics after this restriction.

These data offer various control variables we can utilize in our analyses. From the individual ratings, we are able to see the unique ID of each reviewer, which allows us to control for individual differences between raters. There are 1,389,461 unique reviewers, of whom 131,045 leave multiple ratings, and 48,777 leave ratings for both superhosts and non-superhosts.

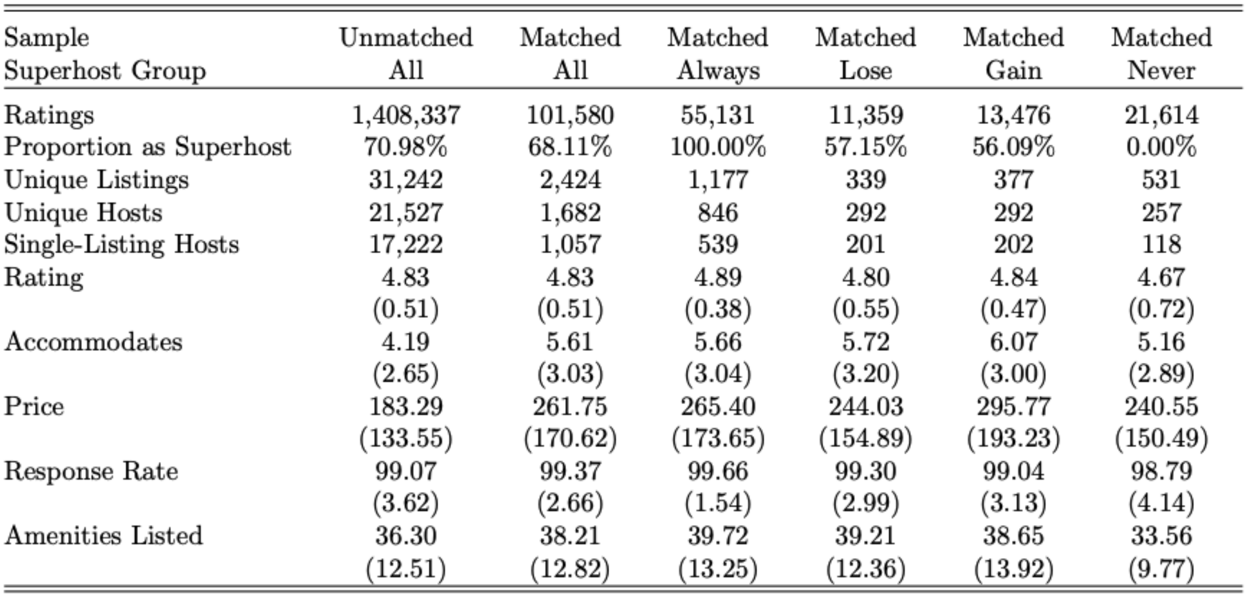
In a given quarter, 14.4% of non-superhosts gain superhost status, while 6.4% of superhosts lose their status in a given quarter. These changes in status are correlated with Airbnb's posted criteria, though imperfectly so—for example, only 65.3% of listings who earn superhost status achieve an average rating of 4.8 or higher in the prior year. From the data we can observe, we would expect more listings that meet all criteria to gain superhost status, and more listings that do not meet all criteria to lose superhost status (Web Appendix C). This is consistent when we only examine single-property hosts. Thus, unobserved variables must be impacting changes in superhost status—including the fact that Airbnb hosts can petition the platform to retain status.

Vrbo Listing Snapshots and Ratings. Many properties that are listed on Airbnb are also listed on Vrbo, a competitor in the peer-to-peer homesharing market. Vrbo operates in a similar fashion as Airbnb—hosts provide information about their listings, and consumers rate their experiences on 1–5 star scales—but superhost status is not present on Vrbo, as this is a certification awarded by Airbnb. Thus, it is possible to estimate the effect of superhost status on Airbnb over time by comparing listings to themselves on Vrbo.

To this end, we collected listing information and ratings from Vrbo in the locations for which we have Airbnb data. Because there is no common key between platforms and both platforms mask listings' locations, we developed a simple algorithm to match Airbnb listings to the corresponding listing for the same property on Vrbo (Web Appendix D). From this process, we were able to match 2,424 unique listings between Airbnb and Vrbo. These matches correspond to 103,987 individual Airbnb ratings and 23,283 Vrbo ratings.

TABLE 3

DESCRIPTIVES OF AIRBNB LISTINGS MATCHED TO VRBO



Sample Superhost Group	Unmatched All	Matched All	Matched Always	Matched Lose	Matched Gain	Matched Never
Ratings	1,408,337	101,580	55,131	11,359	13,476	21,614
Proportion as Superhost	70.98%	68.11%	100.00%	57.15%	56.09%	0.00%
Unique Listings	31,242	2,424	1,177	339	377	531
Unique Hosts	21,527	1,682	846	292	292	257
Single-Listing Hosts	17,222	1,057	539	201	202	118
Rating	4.83	4.83	4.89	4.80	4.84	4.67
	(0.51)	(0.51)	(0.38)	(0.55)	(0.47)	(0.72)
Accommodates	4.19	5.61	5.66	5.72	6.07	5.16
	(2.65)	(3.03)	(3.04)	(3.20)	(3.00)	(2.89)
Price	183.29	261.75	265.40	244.03	295.77	240.55
	(133.55)	(170.62)	(173.65)	(154.89)	(193.23)	(150.49)
Response Rate	99.07	99.37	99.66	99.30	99.04	98.79
	(3.62)	(2.66)	(1.54)	(2.99)	(3.13)	(4.14)
Amenities Listed	36.30	38.21	39.72	39.21	38.65	33.56
	(12.51)	(12.82)	(13.25)	(12.36)	(13.92)	(9.77)

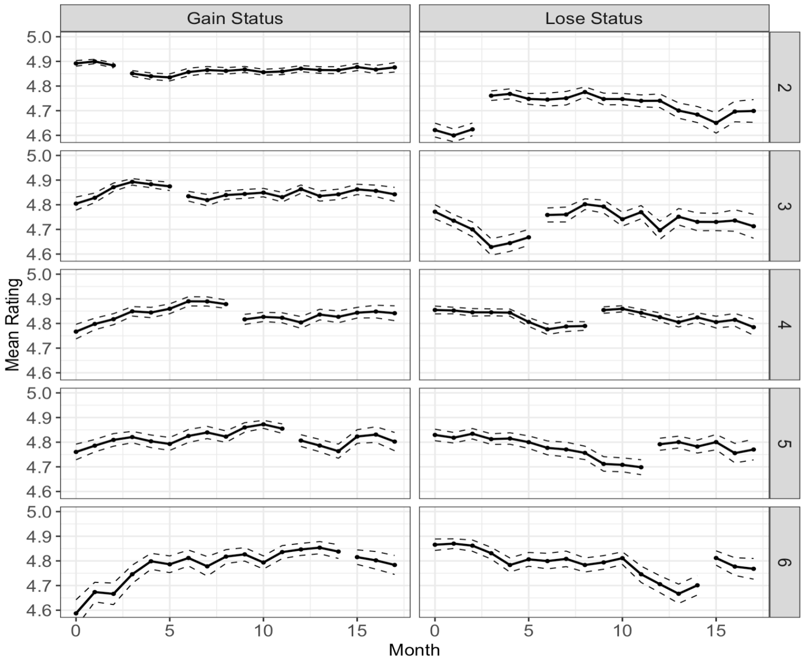
NOTE.— This table includes Airbnb information only.

Table 3 presents descriptive Airbnb statistics for the subset of Airbnb listings we were able to match on Vrbo. The first column includes the subset we were not able to match as a comparison. Samples are quite similar overall; the average Airbnb rating in each group is close across samples, as is the response rate and number of amenities listed. Meanwhile, differences in prices and guests accommodated are likely due to the fact that Vrbo only allows listings for entire homes, while Airbnb allows hosts to rent out a room while they are present.[5]

Model-Free Evidence

Figure 1 plots the average monthly ratings for properties that gain (left column) and lose (right column) superhost status. The visualization is further broken down by the quarter at which the change in status occurs: The top row of panels feature data from listings where the superhost status changed at the first quarter of observation, the second row shows those where the superhost status changed in the second quarter, and so on.

FIGURE 1
MONTHLY AVERAGE RATINGS FOR AIRBNB LISTINGS



NOTE.— Ratings are aggregated at the month level in this plot to add granularity. Superhost status can only change quarterly. The gap in lines within panels represents the time of change. Dashed lines represent upper and lower bounds of 95% confidence intervals.

[5] We do compare information such as prices across platforms. This is because we do not have repeated snapshots of Vrbo listings, and our Vrbo snapshots are from later than our Airbnb snapshots. This does not present a problem for our matching algorithm, as it uses information less likely to change over time (number of accommodated guests, names, host names, descriptions, locations).

As evident in Figure 1, the transitions into and out of superhost status feature changes in ratings consistent with our predictions: When a property gains superhost status, the ratings tend to go down. When a property loses superhost status, the ratings tend to go up. Moreover, it does not appear that listings who gain (lose) superhost status do so because of a short period of abnormally high (low) ratings, which would heighten concerns about "regression-to-the-mean". Specifically, regression to the mean would predict that listings earn (lose) superhost status after a period of abnormally high (low) ratings. In contrast, we observe a pattern of steady increases in ratings followed by a sudden decrease in ratings after obtaining superhost status, whereas a sudden decrease is exactly what would be expected from our hypothesis.[6] Our three identification strategies seek to estimate the causal effect illustrated by Figure 1.

Identification Strategies

Difference-in-Differences in Airbnb Ratings. Our first identification strategy investigates the effect of changing superhost status using a difference-in-differences approach. We first create two subsets of data based on the superhost status of the property at the time of our first observation. Then, within each subset, we run a separate event study comparing those whose superhost status changes to those whose status does not change. Specifically, in one event study we compare those who gain superhost status to those who are never superhosts (Equation 1) and in the other we compare those who lose superhost status to those who are always superhosts (Equation 2). These two models are represented by the following equations:

$$Rating_{iq} = \alpha_1 Gain_i + \alpha_2 Post_q + \delta Gain_i \times Post_q + \beta X_{iq} + \varepsilon_{iq} \qquad (1)$$

[6] We further investigate this regression to the mean concern in Web Appendix E, where we repeat this plot for the listings with low pre-change standard deviations in ratings, finding similar patterns.

$$Rating_{iq} = \alpha_1 Lose_i + \alpha_2 Post_q + \delta Lose_i \times Post_q + \beta X_{iq} + \varepsilon_{iq} \qquad (2)$$

In both, the δ coefficient measures the difference-in-differences—the difference in ratings for listings that gain (in Equation 1) or lose (in Equation 2) superhost status, compared to the difference in ratings at the same time for listings who are never (in Equation 1) or always (in Equation 2) superhosts. Listings can change superhost status in any of quarters 2–6.[7] We analyze the effect of the first change in superhost status only, removing observations once a listing has changed status twice.[8] Therefore, both equations are examples of "staggered difference-in-differences", where treatment is determined at different times for different units. To handle this, we adjust all treated observations such that superhost status changes at $q = 0$. Conceptually, this is as if we shifted the x-positions of lines in Figure 1 such that gaps in all of the panels were vertically aligned.

Because there is no obvious $q = 0$ point for listings who do not change status, we follow the estimation procedure outlined in Callaway and Sant'Anna (2021), which treats each possible treatment timing as a separate event studies—comparing those whose superhost status changes at that time to the entire control group. We estimate these models using the *did* package (version 2.1.2; Callaway and Sant'Anna 2018) in the R programming language (version 4.4.1; R Core Team 2024), clustering standard errors by listing. This procedure estimates δ (the difference-in-differences) as the average treatment effect on treated listings (*ATT*). We report this ATT in-text, as it quantifies the average effect on ratings of gaining or losing superhost status for listings who change status.

[7] This is because we do not observe superhost status prior to quarter 1.

[8] Analyses removing all observations from those who change status more than once are similar (Web Appendix F).

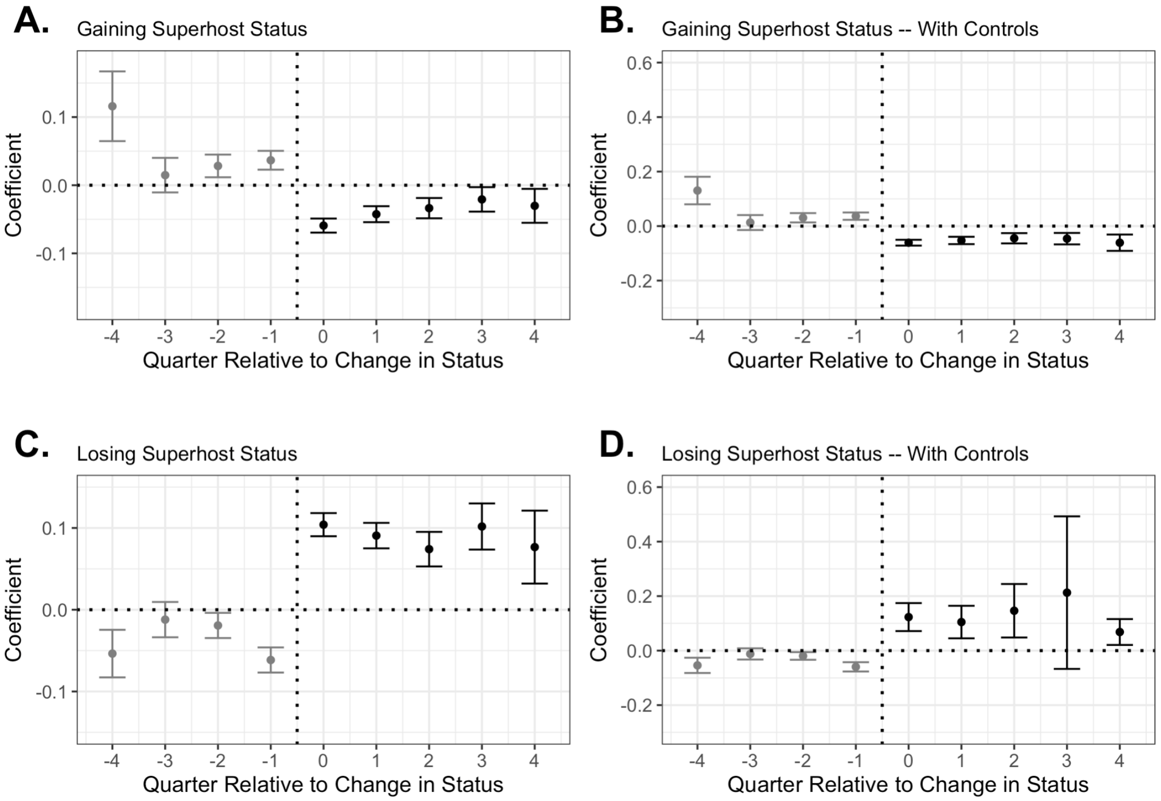
This method allows us to test a key assumption underlying difference-in-differences, which is of parallel pre-trends—that ratings for treated and control listings are parallel prior to the change in status. We do so by calculating an average treatment effect for each quarter relative to treatment across all event studies, expressed by the following equations:

$$Rating_{iq} = \sum_{t=-4}^{4} \delta_t Gain_i \times (q = t) + \varepsilon_{iq} \qquad (3)$$

$$Rating_{iq} = \sum_{t=-4}^{4} \delta_t Lose_i \times (q = t) + \varepsilon_{iq} \qquad (4)$$

We present results for Equation 3 in Figure 2A and Equation 4 in Figure 2C. Each point estimate is δ_t, the estimated ATT between treated and control listings. The black points in each plot demonstrates this treatment effect after treatment, and is consistent with H_1: After changing status, ratings for those who gain status drop compared to non-superhosts, and ratings for those who lose status increase relatively to those who are always superhosts. However, the grey points, which indicate $\delta_t$before changing status, show violation of the parallel pre-trends assumption. In Figure 2A, there is a difference between treated listings (those who gain superhost status) and control listings (never superhosts) prior to treatment (grey points), such that those who gain status see their ratings increase more strongly prior to the change in status. Figure 2C shows the opposite—those who lose status see their ratings decrease relative to those who are always superhosts. If we control for hosts' number of listings, observed response rate, number of ratings, and the listing's price, the pre-trends become less divergent, but still not parallel. This can be seen in Figures 2B and 2D, where the coefficients are closer to—but still significantly different than—zero prior to treatment.

FIGURE 2
AVERAGE TREATMENT EFFECTS OF SUPERHOST STATUS ACROSS TIME PERIODS



NOTE.— Bars represent 95% CI. Panels A and C result from models not controlling for listing attributes. Panels B and D results control for price, hosts' response rate, hosts' number of listings, and the number of ratings received in a quarter.

While the lack of parallel trends is problematic from a causal inference perspective, it should not be surprising in this particular setting: Superhost status is not randomly determined, but earned by hosts. In similar cases, it is common to investigate a specific subset of data where treatment (changing superhost status) is "as good as random." For example, rather than comparing all listings who change status to all listings that never change, we could compare only listings who *barely* changed to listings who *nearly* did. The assumption in such scenarios is that, while treatment is largely not random, the difference between barely and nearly being treated is effectively so. Unfortunately, the impact of unobservables on superhost determination makes this comparison untenable (see Web Appendix C). Instead, we propose an alternative identification strategy, which uses each listing as its own control.

Within-Listing on Airbnb. Our second identification strategy is to take all properties that have periods of both superhost and non-superhost status and compare the ratings they receive during each period, as expressed in Equation 5:

$$Rating_{iqj} = \alpha_1 Superhost_{iqj} + \beta X_{iqj} + \varepsilon_{iqj} \qquad (5)$$

where $Superhost_{iqj}$ is a dummy code indicating superhost status for listing i at quarter q, reviewed by reviewer j. X_{iqj} is a vector of fixed effects for listing, quarter, and reviewer. In this analysis, between-listing and between-time period variation in ratings are removed from each observation via fixed effects. Thus, the effect of superhost status estimated by α_1 is estimated after controlling for all time-invariant features of listings (e.g., location, cleanliness, host attributes, etc.), and any differences between time periods across listings. Listing fixed effects (denoted by *i*) are necessary to identify a causal effect of superhost status on ratings. Without listing fixed effects, the estimated effect of superhost status on ratings would be confounded with differences in the true quality of listings who do and do not earn superhost status. Consistent with this notion, there is a positive estimate of superhost status on ratings when we regress ratings on status without any controls ($\beta_{\text{Superhost}}$ = .180, *t*(1,557,978) = 77.058, *p* < .001).

The inclusion of reviewer fixed effects (denoted by j) allows us to control for an additional concern: the selection of consumers into different listings. If some consumers prefer to stay with superhosts and are more negative in general, listings may receive lower ratings when designated as superhosts not due to increased expectations, but due to a shift in clientele. By including reviewer fixed effects, we are able to remove between-reviewer differences in ratings.

The repeated snapshots from InsideAirbnb also allow us to observe how properties change over time in attributes other than superhost status. For example, we are able to observe the amenities listed, price offered, and number of listings hosts operate for each property in each quarter, allowing us to identify changes in properties and the possibility that hosts become overextended with superhost status, among other changes. Together, this information begins to address the threats to causality presented by time-varying attributes. We perform a specification curve analysis (Simonsohn, Simmons, and Nelson 2020) by running 3,840 variants of Equation 5 to determine if including/excluding any of these potential control variables has a systematic effect on our estimate of α_1. This analysis finds no impact of observable time-varying attributes on the key result and is presented in Web Appendix F.[9]

While this specification curve begins to address time-variant quality by controlling for attributes that we can observe (e.g., demand, price, hosts' response rate, amenities, etc.), it is ultimately limited to attributes we can observe. Furthermore, the fixed-effect regression expressed by Equation 5 does not allow us to test parallel trends. Thus, we employ a third identification strategy, which compares listings on Airbnb to themselves on Vrbo.

Difference-in-Differences Between Airbnb and Vrbo. Many of the properties we observe on Airbnb are also listed on Vrbo. And, while Vrbo customers rate listings in a very similar way to Airbnb customers, they do not see information provided by Airbnb itself—critically, superhost status is absent on Vrbo. If the quality of a listing changes during periods of superhost status (on Airbnb.com), we should see commensurate decreases in ratings on Vrbo during these periods as

[9] Web Appendix F also includes a specification curve for the same models using the sentiment of text reviews as the outcome, finding similar results.

well. However, if the change in ratings is driven by the change in context/expectations as we propose, we should not see any differences in ratings on Vrbo across periods of superhost (vs non-superhost) status, because the superhost tag is part of Airbnb and would not be shown to Vrbo users.

To test this, we again separate listings into those who gained or lost superhost status in our data. Consistent with the identification, we limit observations to a listings' first change in superhost status. We then estimate the following model separately for those subsets:

$$Rating_{iq} = \alpha_1 Airbnb_i + \alpha_2 Post_q + \delta Airbnb_i \times Post_q + \beta X_{iq} + \varepsilon_{iq} \qquad (6)$$

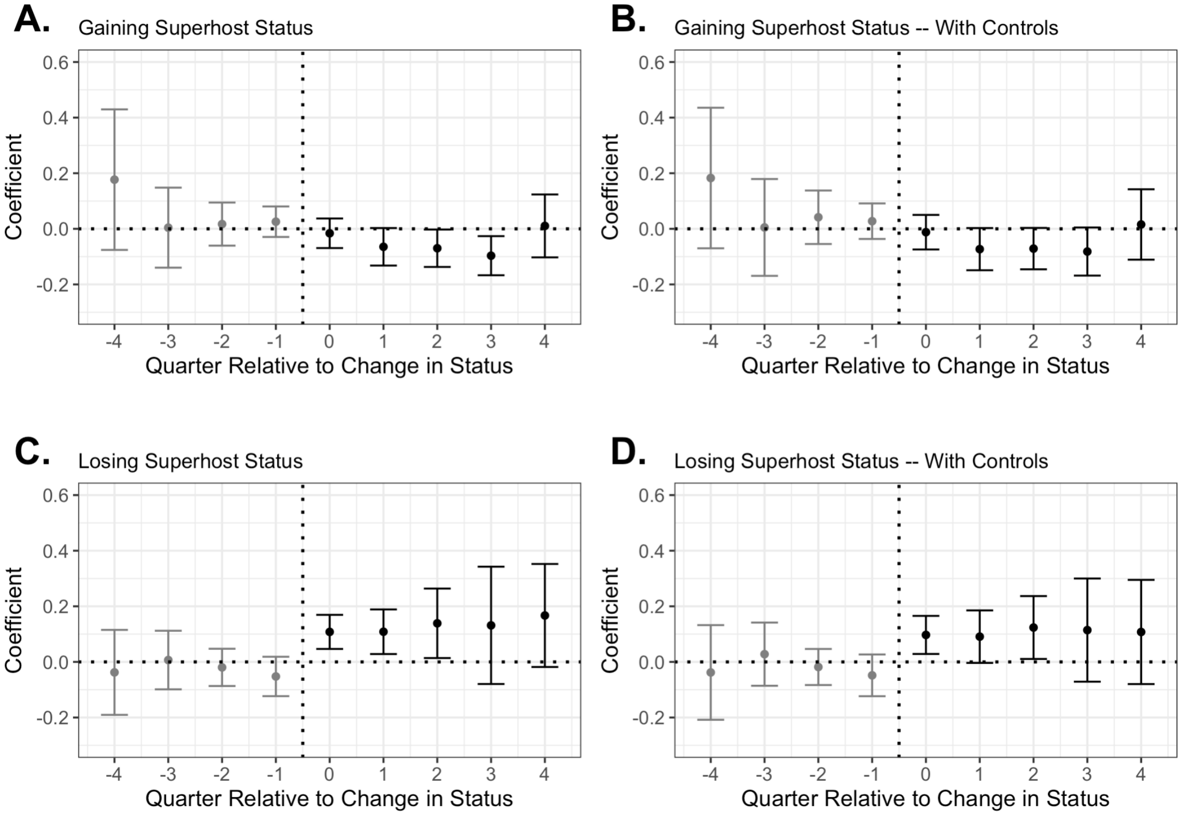
where $Airbnb_i$ is a dummy code indicating whether the platform a rating was provided on was Airbnb, $Post_q$ is a dummy code indicating whether a rating was provided after superhost status changed, and X_{iq} is a vector of listing fixed effects for listing and time period fixed effects. The δ coefficient again measures the average treatment effect on treated listings (ATT); in this model, the ATT quantifies the difference in ratings on Airbnb for listings that gain or lose superhost status *after* changing status, compared to the difference in ratings at the same time experienced for the same listings on Vrbo.

To interpret this coefficient as the causal effect of changing superhost status, we must satisfy similar assumptions as we describe in the discussion of our first identification strategy. In this case, however, we must examine the parallel pre-trends assumption by comparing ratings on Airbnb to those on Vrbo prior to changing superhost status. We investigate these trends with the same Callaway and Santa'Anna (2021) estimation procedure as in our first identification. We test for parallel pre-trends using the following equation, which we estimate separately for those who gain and lose superhost status:

$$Rating_{iq} = \sum_{t=-4}^{4} \delta_t Airbnb_i \times (q = t) + \varepsilon_{iq} \qquad (7)$$

FIGURE 3

AVERAGE TREATMENT EFFECTS OF SUPERHOST STATUS ACROSS TIME PERIODS



NOTE.— Panels A and C result from models not controlling for listing attributes. B and D control for price, host response rate, host number of listings, and ratings received in a quarter.

These data largely satisfy the parallel trends assumption (Figure 3). In all four models (comparing gainers to never superhosts and losers to always superhosts, with and without controls), all pre-change coefficients are not different from zero. This means that there is no individual quarter where the change in ratings is significantly different between groups. In addition, we do not see substantial differences in the last quarter prior to treatment. This reduces the potential concern that listings have abnormal periods immediately before changing status, which would heighten potential concerns about “regression-to-the-mean.”

Specifically, regression to the mean should not produce parallel pre-trends in the Vrbo data. The core idea behind a regression-to-the mean based explanation is that superhost status is acquired after a period of abnormally high ratings and—importantly—that these ratings are abnormally high due to random variability in perceptions or performance. There is no reason why the variability on Airbnb should correspond with identical variability on Vrbo. In other words, regression to the mean should not produce a systematic pattern in the Vrbo data: Before a property becomes a superhost on Airbnb, you should see a rise in the ratings on Airbnb (because superhost status is cause by that rise in ratings), but the pattern of data on Vrbo should be—in expectation—flat. This would lead to systematically different pre-trends, which we do not observe

Results

Difference-in-Differences in Airbnb Ratings. Our first set of results are those of the difference-in-differences models of Equation 1 and Equation 2. These models compare the change in ratings after changing superhost status for listings who gain status to those who are never superhosts (Equation 1) and for listings who lose status to those who are always superhosts (Equation 2). As with our test of pre-trends, we estimated these models using the *did* package (version 2.1.2 Callaway and Sant'Anna 2018) in the R programming language (version 4.4.1; R Core Team 2024), which estimates the ATT—the average treatment effect on treated listings (i.e., the effect a change in status has on ratings for listings that change status). To account for interdependence, we cluster standard errors by listing.

Results support H_1 and are presented in Table 4 both with and without controls. Listings who gain superhost status see a more substantial decrease in ratings after gaining relative to those who are never superhosts in the same time periods (ATT = –.037, SE = .004, 95% CI = [–.046, –.029]; Model 1). This ATT suggests that the average rating for a listing is .037 stars lower on average after gaining superhost status than it would have been if the listing remained a non-superhost, an effect which is equivalent to 8.3% of the standard deviation in ratings for listings who gain status (*SD* = .444). This effect is consistent after controlling for price, the number of reviews received in that quarter, hosts' response rate, and hosts' number of listings (Model 2), which could plausibly correlate with any time-varying changes in quality.

TABLE 4
RESULTS OF DIFFERENCE-IN-DIFFERENCES FOR GAINING AND LOSING SUPERHOST STATUS

Model	1	2	3	4
Treated Group	Gain	Gain	Lose	Lose
Control Group	Never	Never	Always	Always
ATT	$-.037^{***}$	$-.053^{***}$	$.089^{***}$	$.131^{***}$
SE	(.004)	(.006)	(.006)	(.034)
Controls		✓		✓
Listing FEs	✓	✓	✓	✓
Quarter FEs	✓	✓	✓	✓
Observations	469,752	429,351	1,040,165	1,009,031
Mean DV	4.72	4.72	4.88	4.88

NOTE.– Controls are price, hosts' response rate, hosts' number of listings, and the number of ratings received in a quarter.

Models 3 and 4 replicate Models 1 and 2 for listings who lose superhost status, comparing them to listings who are always superhosts. Those who lose status see a more substantial increase in ratings after losing (ATT = .089, SE = .006, 95% CI = [.078, .101]), which represents 14.7% of the standard deviation in ratings for these listings (*SD* = .613). This is also consistent after controlling for observable attributes (Model 4). As discussed above however, this identification violates two key assumptions of difference-in-differences required for causal inference: (i) units can influence their treatment status and (ii) we do not observe parallel trends prior to changing status. Thus, we suggest interpreting the results from this first analysis with caution.

Within-Listing on Airbnb. Our second analysis strategy controls for between-listing differences by analyzing variation in ratings entirely within listings. This strategy follows Equation 5, which also allows us to control for between-quarter and between-reviewer differences in ratings. In Table 5, we present the results for five specifications of this model. We cluster standard errors at the listing level in each.

Model 1 controls only for time-invariant differences in quality between listings by estimating a fixed-effect for listings, removing between-listing variation in ratings. With this specification, we find that ratings for listings are lower when listings are superhosts than when they are not ($\beta_{\text{Superhost}} = -.041$, $t(1{,}524{,}305) = -20.779$, $p < .001$, median within-unit Cohen's $d = -.139$). While .04 stars may appear like a small effect in isolation, properties with variation in superhost status have an average yearly rating of 4.8/5, so .04 corresponds to 20% of the gap between the average rating and the scale maximum.

TABLE 5
RESULTS FOR WITHIN-LISTING AIRBNB MODELS

Model	1	2	3	4	5
Superhost	$-.041^{***}$	$-.023^{***}$	$-.041^{***}$	$-.022^{***}$	$-.022^{***}$
	(.002)	(.004)	(.002)	(.004)	(.004)
N Ratings					$.009^{**}$
					(.003)
Listing FEs	✓	✓	✓	✓	✓
Quarter FEs			✓	✓	✓
Reviewer FEs		✓		✓	✓
Observations	1,557,980	1,557,980	1,557,980	1,557,980	1,557,980
R^2	.094	.942	.094	.942	.942
Adj. R^2	.074	.344	.074	.344	.344
Residual SE	.491	.413	.491	.413	.413
df	1,524,305	137,009	1,524,400	137,004	137,003

NOTE.— The number of ratings is log transformed.

Models 2–5 then demonstrate the robustness of this result to the inclusion of reviewer and quarter fixed effects. All four models yield a negative effect of superhost status, meaning that the same listing received worse ratings during the period(s) it was a superhost. Model 5 additionally addresses a potential concern that Airbnb hosts are unable to provide the same service after attaining superhost status. Because superhost status leads to an increase in the number of ratings a listing receives in a quarter ($\beta_{\text{Superhost}}$ = .248, t(152,581) = 6.906, p < .001), one could wonder if hosts are overwhelmed when superhosts, leading to lower ratings. This suggestion is not supported in Model 5, where the effect of superhost status remains negative and significant.

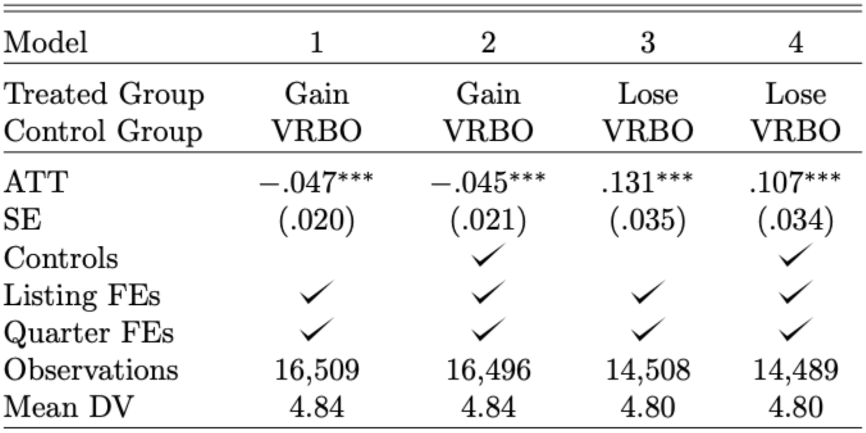
In Web Appendix F, we extend the breadth of models we can test beyond what is possible to communicate in a table. Specifically, we present a specification curve analysis (Simonsohn, Simmons, and Nelson 2020), in which we investigate the coefficient of superhost status on ratings from 3,648 variants of the focal model. Each variant specification is a unique combination of choices of (i) data, (ii) control variables, (iii) fixed-effects, and (iv) standard error clustering we consider to be reasonable variations of our main model. From 3,600 models (98.7% of all models), we find a negative estimated effect of superhost status, which is statistically significant in 2,253 (61.8%). The estimate is significant and negative in all models that do not include a reviewer fixed effect is only positive in models that include a reviewer fixed effect and only consider a subset of the total data. The median coefficient estimate is –.025 overall, –.017 for models with reviewer fixed effects, and –.050 for models without. We also replicate this specification curve analysis using the sentiment in the text of each review as our dependent variable, finding similar results.

Finally, in Web Appendix F, we also investigate the heterogeneity of the estimated effect of superhost status across different types of listings. We do so by augmenting Equation 5 by including listing attributes—including an indicator for single-listing hosts—and their interaction with superhost status in the model. We find that the effect is largely stable between different listings, with two noteworthy exceptions. First is that listings from hosts with multiple listings show a less negative effect of superhost status on ratings ($\beta_{Superhost \times Multi}$ = .028, t(1,524,298) = 6.771, p < .001, 95% CI = [.020, .036]). Second is that superhost status has a less negative effect for higher-priced listings ($\beta_{Superhost \times logPrice}$ = .007, t(1,523,862) = 2.306, p = .021, 95% CI = [0.001, .013]). In other words, higher-priced listings have smaller changes in ratings between periods with and without superhost status. This could be due to a crowding-out of the effect of superhost status on expectations, as price is also known to heighten expectations.

Airbnb-Vrbo Difference-in-Differences. Our final set of analyses test H_1 by estimating Equation 6 separately among those who gain and lose superhost status. As with our test of pre-trends, we estimated these models using the *did* package (version 2.1.2; Callaway and Sant'Anna 2018) in the R programming language (version 4.4.1; R Core Team 2024). To account for interdependence in ratings for listings, we cluster standard errors at the listing level.

Results support H_1. Specifically, listings who gain superhost status see a more substantial decrease in ratings on Airbnb after gaining than they do on Vrbo (Table 6 Model 1; ATT = –.047, SE = .021, 95% CI = [–.089, –.005], representing 10.2% of the standard deviation in ratings for these listings). And listings who lose superhost status see a more substantial increase in ratings on Airbnb after losing relative to themselves on Vrbo (Table 6, Model 3; ATT = .131, SE = .035, 95% CI = [.058, .204], representing 23.3% of the standard deviation in ratings for these listings). We also replicated these results in models that controls for observable listing attributes (Models 2 and 4). Because we do not have historical Vrbo snapshots, all information for these controls comes from Airbnb. Finally, we assess the heterogeneity of this effect in Web Appendix G, finding that these results are consistent across subpopulations, though larger for listings from hosts with only a single listing.

TABLE 6
RESULTS OF DIFFERENCE-IN-DIFFERENCES BETWEEN AIRBNB AND VRBO RATINGS FOR GAINING AND LOSING SUPERHOST STATUS



Model	1	2	3	4
Treated Group	Gain	Gain	Lose	Lose
Control Group	VRBO	VRBO	VRBO	VRBO
ATT	−.047***	−.045***	.131***	.107***
SE	(.020)	(.021)	(.035)	(.034)
Controls		✓		✓
Listing FEs	✓	✓	✓	✓
Quarter FEs	✓	✓	✓	✓
Observations	16,509	16,496	14,508	14,489
Mean DV	4.84	4.84	4.80	4.80

NOTE.– Controls are price, hosts' response rate, hosts' number of listings, and the number of ratings received in a quarter.

Discussion

Across all three identification strategies, the analyses are consistent with the claim that the superhost designation leads to properties receiving lower ratings. The first—utilizing a traditional difference-in-differences between groups of listings—is perhaps easiest to understand and visualize, and allows us to investigate trends over time directly. However, we note two potential concerns with this strategy: listings can influence their treatment status and pre-trends are not parallel.

The second identification strategy does not estimate a time trend, but removes variation between listings, time periods, and reviewers in ratings through fixed-effects. Thus, each listing acts as its own control, assuaging the concern about creating control groups of listings. Notably, results from models in this strategy that include reviewer fixed-effects show much weaker effects of superhost status on ratings ($\beta_{\text{Superhost}}$ = –.022 with reviewer fixed-effects, $\beta_{\text{Superhost}}$ = –.041 without). This suggests that at least some portion of the estimated effect of superhost status is attributable to selection of consumers into superhost vs. non-superhost listings.

We also find that the effect of superhost status on ratings is remarkably robust across model specifications and types of listings (Web Appendix F). In Web Appendix F, we test a series of interactions of listing characteristics (e.g., price, number of listed attributes) with superhost status in Equation 5. Results of these models demonstrate consistent effects across listings, consistent with our H_1., which centers only on the cognitive impact of superhost status in changing the comparison raters bring to mind when creating ratings. Two exceptions to this consistency are hosts with multiple listings and high-priced listings, who see a smaller effect of superhost status. We note that the latter interaction is also consistent with H_1. Prices also affect expectations (Cadotte et al., 1983), so higher-priced listings are likely already evaluated in a similar way as superhost-certified listings. This may effectively crowd out the effect of superhost status.

Finally, our third strategy resolves the issue of time-variant unobservables by comparing listings who gain or lose superhost status against themselves on Vrbo—a similar platform to Airbnb, where the superhost designation is not presented. In these models, losing superhost status appears to have a much stronger impact on increasing ratings (Table 6, Model 4; *ATT* = .107) than gaining status has on decreasing ratings (Table 6, Model 2; *ATT* = –.045). Though we predict the difference in direction of these results—that losing status increases ratings, and gaining status decreases them—we did not predict a difference in the size of these effects. This is because consumers only see a listing's current superhost status, not a listing's former status.

Consumers cannot see this time trend in superhost status. While this difference cannot result from consumers reacting to gaining vs losing status differently, it is consistent with other research on ratings. For example, Godes and Silva (2012) find that ratings for books decrease over time. Thus, the decrease in ratings on Airbnb after gaining status may be partly hidden by the effect of time, while those who lose status see a larger positive effect because their Airbnb ratings buck this trend. Regardless of the cause, this difference has important economic implications for hosts. Specifically, it suggests that the negative effect of losing superhost status on demand is softened over time by an increase in ratings.

The analyses above attempt to go beyond establishing a descriptive result (i.e., properties receive lower ratings when they have superhost status) to identify the *causal* effect of superhost status on ratings (i.e., superhost status *causes* properties to receive lower ratings). We note a possible challenge with this latter interpretation: User-generated ratings (our dependent measure) and superhost status (our independent measure) are inherently endogenous. Our analyses attempt to circumvent possible issues with this endogeneity. For example, by looking within property, we can address concerns about property quality serving as a common cause for both variables. We are further able to assess the robustness of these within-listing models in Web Appendix F, finding no reason for concern about our conclusion for H_1.

An additional concern is regression-to-the-mean—that listings earn superhost status after periods of abnormally high ratings, thus demonstrating lower ratings while superhosts not due to our hypothesis, but due to their ratings returning to “true” quality. While this is difficult to directly rule out in our context, we note that it is inconsistent with the parallel pre-trends we observe between Airbnb and Vrbo, as we would not expect noise in ratings to be consistent across platforms. Other analyses included in Web Appendix E begin to address this concern empirically. Ultimately however, these remaining concerns motivate Study 2A, in which we employ an experimental approach that breaks the endogeneity between superhost status and ratings and allows for unambiguous causal inference.

EXPERIMENTAL EVIDENCE

We corroborate and extend our conclusions from Study 1 in three follow-up studies, described in brief here. First, Study 2A provide an experimental replication of Study 1. We find that a property receives lower ratings when it is a superhost (vs when it is not; H_1). We note that this result is inconsistent with regression-to-the-mean and other potential concerns from Study 1 relating to an endogenous treatment. Next, two studies assess H_2, which is the prediction that consumers do not anticipate the effect of quality signals on ratings when makings choices using those ratings. In each study, we assess the competing influences of superhost status when consumers chose between listings: Being a superhost should provide a positive quality signal to prospective consumers, but should be accompanied by the potential negative effect (H_1; documented in previous studies) of reduced ratings. Study 2B utilizes the ratings provided by Study 2A participants as stimuli, while Study 3 utilizes ratings from real Airbnb listings. In both studies we find support for H_2: Prospective consumers are insufficiently sensitive to the effect of superhost status on ratings and instead chose as if ratings are an unbiased proxy for quality.

All lab studies were pre-registered, and any deviations from the pre-registrations have been noted in text. All code, data, materials, and pre-registrations are available on our OSF repository (https://osf.io/3he6c/?view_only=e031a89ca6fd464ebb67de90e0363014).

STUDY 2: AIRBNB EXPERIMENT

Study 2 involves two parts. In Part A, we ask participants to rate a hypothetical stay at an Airbnb. We experimentally manipulate whether the property the participant rates was described as a superhost (or not) and assess whether this has an effect on ratings (H_1). In Part B, we ask prospective consumers to choose between two properties: one is a superhost and the other is not. To reinforce the connection between our two hypotheses—the population of consumers who are influenced by certifications when creating ratings is the same population that underappreciates this influence when using ratings—we present the ratings provided by participants in Part A as stimuli in Part B. This allows us to test assess how prospective consumers incorporate the joint effects of a quality-signaling certification (superhost status) and the reduced ratings that certification entails (H_2).

Study 2A: Generating Ratings

Participants & Procedure

Five-hundred sixty-two workers from Amazon Mechanical Turk (AMT) started Study 2A.[10] Following our preregistration, 59 participants were removed for failing an attention check. One additional participant passed the attention check, but did not respond to the dependent measure, yielding a final sample of 502 participants. These participants were randomly assigned

[10] There was a mistake in the Qualtrics when we first ran this study. We tested our stimuli in a pilot study, varying whether the "story" about the Airbnb was "positive" or "negative". Results from the pilot suggested using only the positive condition, which we pre-registered. However, we did not remove the "negative" condition at launch. Therefore, there are 167 observations who saw a "negative" story. We remove them. We do not remove those who saw the "positive" story at the same time.

to one condition in a 2 (superhost status: yes vs. no) × 2 (stimulus set: A vs. B) between-subject design: This varied the superhost status of a prospective Airbnb and the set of property pictures used as stimuli.

All participants were told to imagine they were thinking about taking a vacation in Las Vegas with three friends. Participants were told that they wanted accommodation that would provide a fun, relaxing stay. In the "superhost" condition, participants were told *"After discussing potential apartments with the rest of your group, you chose to stay at the following Airbnb property. This Airbnb property has been designated as a superhost by the platform. You have decided to stay at a superhost."* In the "non-superhost" condition, participants were told *"After discussing potential apartments with the rest of your group, you chose to stay at the following Airbnb property. This Airbnb property has not been designated as a superhost by the platform. You have decided to stay at a property that is not listed as a superhost."*

In both conditions, participants were shown a set of pictures of a real Airbnb property (randomly selected from two possible sets of pictures: set A or set B; Figure 4). Then, participants in both conditions read about having the exact same actual experience at their property: *"Your stay was good, but definitely not great. Specifically, the Airbnb did not have any of those nice little "extras" that make a lot of Airbnbs special. Your host did not provide any recommendations, nor stock the cabinets or fridge, and there was not even cookware in the kitchen."*

After reading about their stay, participants were asked to rate this Airbnb experience on a 1–5 star scale, mirroring how Airbnb's elicits ratings. We predicted that participants would provide higher average ratings in the non-superhost condition, as the inclusion of the superhost tag should elicit an unfavorable frame of reference in participants when creating ratings (H_1).

Analysis & Results

A linear regression found no significant difference in ratings across the two sets of pictures used as stimuli (M_A = 3.11, M_B = 3.04; t(500) = –.970, p = .331), so we collapse across stimulus set condition—consistent with our pre-registration—and analyze the difference in ratings between superhost conditions alone. Consistent with H_1, we found a significant difference between superhost conditions on ratings: The same property received lower ratings when it was designated as a superhost ($M_{Superhost}$ = 2.96) than when it was not ($M_{Non\text{-}Superhost}$ = 3.19; t(500) = –3.109; p = .002; d = –.278).[11]

STUDY 2B: CHOICE

Study 2A found support for H_1 in a controlled experimental setting. In Study 2B, we take the actual ratings from Study 2A and show them to prospective consumers tasked with choosing between two Airbnbs. We decided to use the ratings from Study 2A as stimuli to strengthen the connection between our hypotheses: while consumers are influenced by certifications when creating ratings, the same population of consumers insufficiently appreciates this influence when using ratings. We assess whether these prospective consumers are unduly influenced by these frame-dependent ratings, even when they are aware of—and thus can theoretically correct for—the differences in context. We predicted that participants would be more likely to choose an option with high ratings but without superhost status than an option with lower ratings that has

[11] This was consistent across property conditions, as a secondary analysis predicting rating with superhost condition, property stimuli, and their interaction found no significant interaction between superhost and property conditions on ratings ($M_{SuperhostA}$ = 3.01; $M_{Non\text{-}SuperhostA}$ = 3.20; $M_{SuperhostB}$ = 2.92; $M_{Non\text{-}SuperhostB}$ = 3.18; t(498) = –.491, p = .623).

superhost status. Such a pattern of results would suggest that consumers think star ratings are a good point of comparison, even above other information, and that they neglect the role of expectations in creating ratings.

Participants & Procedure

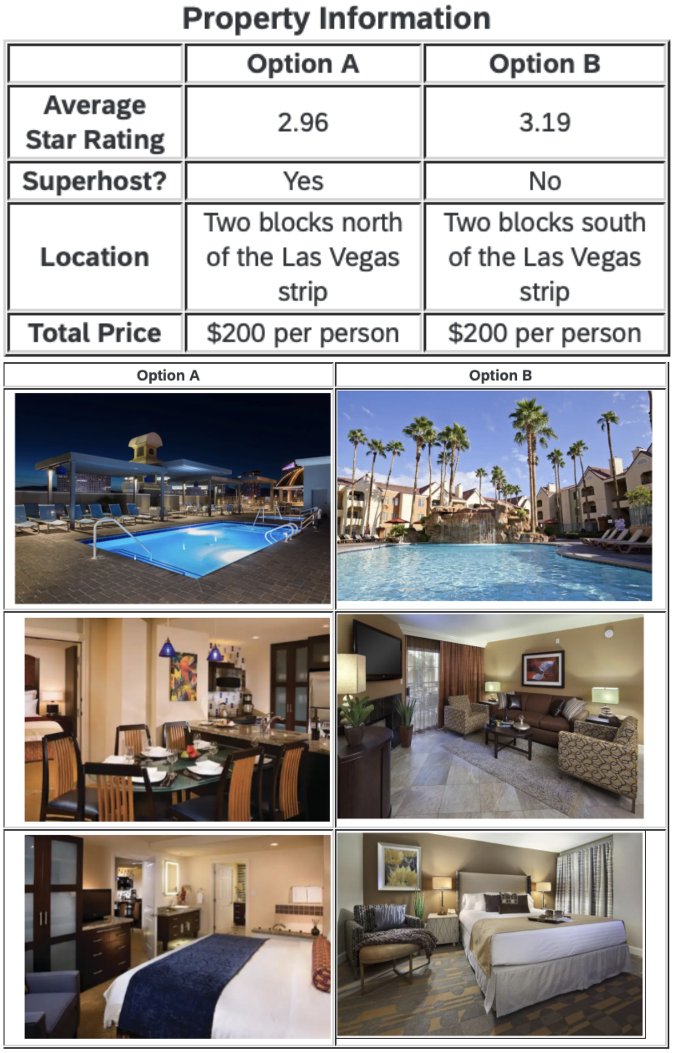
Six-hundred forty-one participants were recruited from AMT. Of these, 42 failed the same attention check from 2A and were removed, leaving us with 599 participants in our sample. All participants were shown the same cover story, which was identical to 2A. However, instead of being told that they had decided to stay at a given property, we asked participants to choose between two.

Participants were shown a table of two available properties. The two properties were based on the attributes used—and average rating received—in the two conditions from Study 2A. For every participant, one property was presented as a superhost with an average rating of 2.96, the other as a non-superhost with an average rating of 3.19. We counterbalanced the order in which the superhost option was displayed (A or B) between participants. Each property was also accompanied by a set of pictures (set A and set B from Study 2A), and additional information about location and price (Figure 4). This information was presented in the same position for every participant, such that property pictures and location information were not confounded with superhost status or ratings.

On the next page, participants responded to our two pre-registered dependent variable measures. First, *"Which of these two Airbnbs do you think is higher quality?"* (0–5, "Definitely A"–"Definitely B"), which we recode such that higher scores correspond to preference for the superhost. Second, *"Which of these two Airbnbs would you choose to stay at?"* ("A", "B", or "No preference"), which we recode such that –1 = non-superhost, 0 = no preference, 1 = superhost. Consistent with H_2, we predicted that participants would indicate that they thought the non-superhost (higher-rated) listing was higher quality. Likewise, we predicted that participants would prefer to stay at the non-superhost (higher-rated) listing.

FIGURE 4

STUDY 2B STIMULI WITH OPTION A AS SUPERHOST



Property Information

	Option A	Option B
Average Star Rating	2.96	3.19
Superhost?	Yes	No
Location	Two blocks north of the Las Vegas strip	Two blocks south of the Las Vegas strip
Total Price	$200 per person	$200 per person

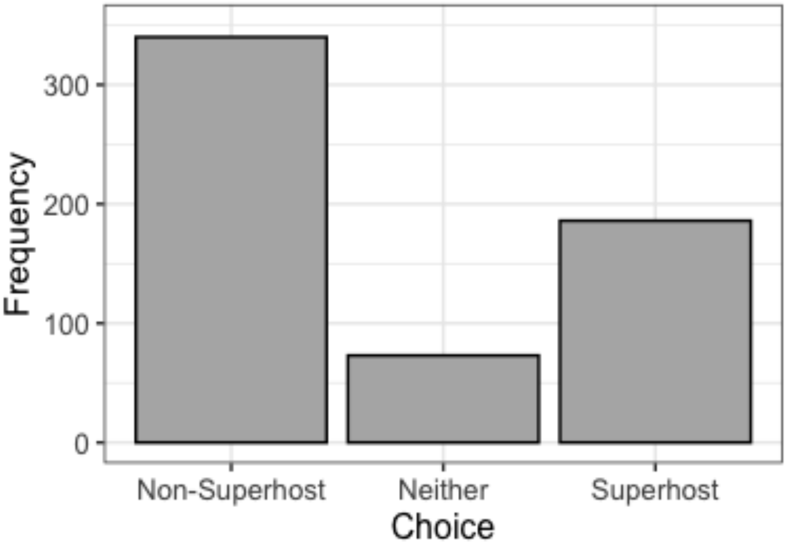
NOTE.–In this example, Option A is described as the superhost with a lower average rating. 49.9% of participants saw this exact stimuli. The other 50.1% saw the average rating and superhost label information swapped, with Option B described as the superhost with a lower average rating. The order of pictures and location information did not vary between participants.

Analysis & Results

For the first dependent measure, we found that quality perceptions differed across the order in which the superhost was presented in the table (A or B; $M_A = 1.98$, $M_B = 2.28$; $t(597) = 3.082$; $p = .002$; $d = .252$). Therefore, we do not collapse across this factor, although results are the same if we do.

Controlling for the order of the superhost listing in the table, the average quality rating is significantly different from the scale midpoint of 2.5 ($M = 2.13$; $t(597) = –7.43$; $p < .001$), with participants thinking the non-superhost (higher-rated) listing is of higher quality.

FIGURE 5
STUDY 2B CHOICES



Consistent results were observed for the choice dependent measure. Controlling for the order the superhost appeared, the average participant was more likely to select the non-superhost (higher-rated) listing ($M = –.26$; $t(597) = –7.58$; $p < .001$). In total, 56.76% of participants chose to stay with the non-superhost (but higher-rated) listing, compared to 31.05% for the superhost, and 12.19% indicating no preference (Figure 5). Therefore, having higher ratings was associated with a 82.8% increase in choice.

Discussion

Study 2 provides support for both of our hypotheses. First, 2A replicates the findings of Study 1 with greater experimental control: Participants gave lower ratings for an experience at a superhost property compared to the exact same experience at a non-superhost property. Second, 2B allows us to test H_2 for the first time. In doing so, we observe evidence suggesting that consumers do not anticipate the effect of mental context on star ratings, even when information about that context is presented directly to them.

This second piece is, in our opinion, key to our manuscript. Researchers and marketers have long talked about the effects of mental context on consumers' evaluations (e.g., Oliver 1980; Parasuraman et al. 1988). Past research even suggests that consumers are aware of this impact of context—often through expectations—on the ratings *they* create (Parasuraman et al. 1985; 1988). Despite this knowledge however, they do not seem to adjust for the potential impact of mental context when using other consumers' ratings to make decisions. If they did, we would not see the results of Study 2B.

We strengthen the support for this finding first in Study 3. This study addresses a potential concern arising from the low ratings observed in Study 2A, which become stimuli in Study 2B. These ratings are quite low, potentially making this comparison unlikely to represent the choices made by real Airbnb customers. Therefore, Study 3 uses real Airbnb stimuli, including ratings.

STUDY 3

Study 3 was designed to test H_2 using stimuli that directly mimic the choices consumers make on Airbnb.com. While 2B benefitted from using ratings from real participants for identical experiences, a relevant concern is that those ratings were too low to be superhosts at all, and that the averages presented straddled 3/5 stars. Therefore, Study 3 does not use participants' ratings. Instead, we present information of two actual Airbnb superhosts that have variation in their average ratings, and randomly drop the superhost designation from one at a time. This allows us to experimentally test the inclusion of superhost status on choice of real Airbnb superhosts.

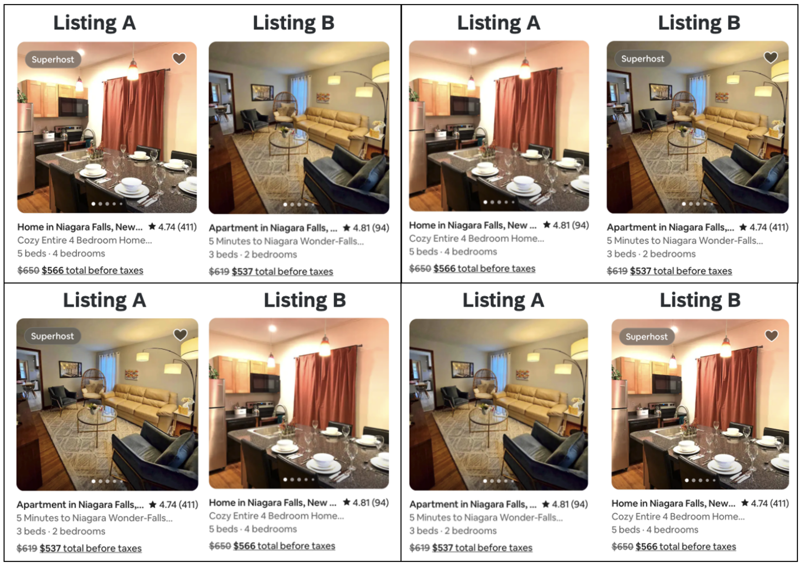
Participants & Procedure

We recruited five hundred participants from CloudResearch by Connect, initially receiving five hundred and five responses. Consistent with our pre-registration, we removed the six observations from three participants whose IP address was duplicated. A further two participants provided no responses. Thus, our final sample is 497 participants. All participants read that they would be shown four sets of two Airbnb properties, and that we would like to know what they thought about the properties, and which they would rather stay at.

Participants were then shown the four sets of properties, one at a time (Figure 6) and in random order. These properties were real Airbnb superhosts in one of four american locations—Los Angeles, Niagara Falls, San Francisco, and Moab. All listings were superhosts at the time of the study, and had similar prices, accommodated similar numbers of guests, but had varying average ratings. We chose to only include superhosts listings to ensure that the ratings we presented to participants were at levels they could realistically observe from superhosts.

Within each location set, we randomized the order in which the two listings were presented to participants. To create variation in superhost status within sets, we randomly dropped the superhost tag from one listing in each set, while to present a conservative test of H_2, we swapped the ratings information, such that the non-superhost always had the higher average rating. Therefore, property information (e.g., price, beds, pictures) and superhost status/ratings were not confounded with each other, nor with the order of presentation. While restricting superhosts to always have lower ratings removes our ability to quantify the effect of platform certifications and ratings on choice independently, that was not the purpose of this study. Instead, the purpose of this study was to investigate whether participants—when presented with divergent information through a platform certification and ratings—would be influenced by the certification, or the ratings. See Figure 6 for an example of stimuli.

FIGURE 6
STIMULI FOR NIAGARA FALLS AIRBNB IN STUDY 3



Listing A
Superhost
Home in Niagara Falls, New... ★ 4.74 (411)
Cozy Entire 4 Bedroom Home...
5 beds · 4 bedrooms
~~$650~~ $566 total before taxes

Listing B
Apartment in Niagara Falls, ... ★ 4.81 (94)
5 Minutes to Niagara Wonder-Falls...
3 beds · 2 bedrooms
~~$619~~ $537 total before taxes

Listing A
Home in Niagara Falls, New ... ★ 4.81 (94)
Cozy Entire 4 Bedroom Home...
5 beds · 4 bedrooms
~~$650~~ $566 total before taxes

Listing B
Superhost
Apartment in Niagara Falls,... ★ 4.74 (411)
5 Minutes to Niagara Wonder-Falls...
3 beds · 2 bedrooms
~~$619~~ $537 total before taxes

Listing A
Superhost
Apartment in Niagara Falls,... ★ 4.74 (411)
5 Minutes to Niagara Wonder-Falls...
3 beds · 2 bedrooms
~~$619~~ $537 total before taxes

Listing B
Home in Niagara Falls, New ... ★ 4.81 (94)
Cozy Entire 4 Bedroom Home...
5 beds · 4 bedrooms
~~$650~~ $566 total before taxes

Listing A
Apartment in Niagara Falls, ... ★ 4.81 (94)
5 Minutes to Niagara Wonder-Falls...
3 beds · 2 bedrooms
~~$619~~ $537 total before taxes

Listing B
Superhost
Home in Niagara Falls, New... ★ 4.74 (411)
Cozy Entire 4 Bedroom Home...
5 beds · 4 bedrooms
~~$650~~ $566 total before taxes

NOTE–. Participants see one of these four options.

On the next page, participants responded to our two pre-registered dependent variable measures. These were the same as in Study 2B. Finally, we asked participants how frequently they stay at Airbnbs when they travel (0–10, never–always; $M = 3.07$, $SD = 2.84$) and how frequently they stay with superhosts when they do stay at Airbnbs (0–10, never–always ; $M = 3.15$, $SD = 3.22$). We did not explain superhost status to participants, but note that Airbnb does not explain superhost status to prospective consumers when presenting listings either.

Analysis & Results

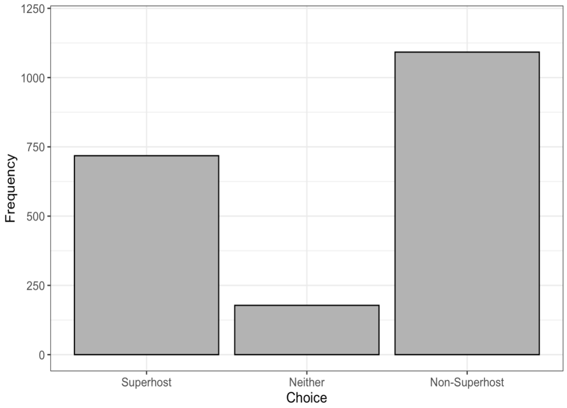
To test each dependent measure, we estimate an intercept-only regression with standard errors clustered by participant.[12] We found that quality perceptions differed across the order in which the superhost was presented (A or B; $M_A = 2.77$, $M_B = 2.9$; $t(1{,}986) = 2.027$; $p = .043$; $d = .088$). Therefore, we do not collapse across this factor, although results are the same if we do (Web Appendix H).

Controlling for the order of the superhost listing in the table, the average quality rating is significantly different from the scale midpoint of 2.5 ($M = 2.83$; $t(1{,}986) = 9.058$; $p < .001$), with participants thinking the non-superhost (higher-rated) listing is of higher quality. To test whether a lack of awareness of superhost status drives this result, we estimated the same regression on the subset of participants who indicated staying at Airbnbs and staying with superhosts more frequently than the median participant (*median* = 2.5 and 2/10, respectively). We observe nearly identical results ($M = 2.83$; $t(790) = 5.399$; $p < .001$).

[12] Note that our pre-registration also included fixed-effects for participant. However, this removed all variation in our dependent measure, so we do not include that fixed-effect.

Consistent results were observed for the choice dependent measure. Controlling for the order the superhost appeared, the average participant was more likely to select the non-superhost (higher-rated) listing (*M* = .19; *t*(1,986) = 8.337; *p* < .001, *d* = .042). In total, 54.93% of participants chose to stay with the non-superhost-tagged listing, compared to 36.12% for the superhost, and 8.95% indicating no preference (Figure 7). Therefore, having higher ratings was associated with a 52.09% increase in choice.

FIGURE 7
STUDY 3 CHOICES



These results are consistent if we treat each city as its own experiment, with the exception of Niagara Falls, which has a non-significant difference in the same direction (Web Appendix H). We also find the same results in the subset of participants who indicated staying at Airbnbs and staying with superhosts more frequently than the median participant (*M* = .15; *t*(790) = 4.053; *p* < .001). Of these, 53.54% chose to stay with the non-superhost (but higher-rated) listing, compared to 38.76% for the superhost, and 7.7% indicating no preference.

Discussion

Study 3 further supports H_2 by demonstrating the influence of ratings on quality perceptions and choice. Specifically, we see that consumers do not properly anticipate the effect of mental context on user-generated ratings.

GENERAL DISCUSSION

Across four studies—three laboratory and one real-world—evidence suggests that platform-created certifications directly affect user-generated ratings, and that prospective consumers underappreciate this possibility. Products and services that are signaled as high quality are judged more harshly by consumers giving ratings. This is problematic because prospective consumers to not anticipate this influence on ratings, diminishing the effectiveness of quality signals in stimulating demand. This is first demonstrated in a large sample of ratings for Airbnb superhosts, where three identification strategies converge to illustrate a negative effect of superhost status on ratings, controlling for objective quality and trends over time. We then test that finding with true randomization in a lab setting, which also allows us to observe consumers' choices in this context. These results are rooted theoretically in the expectation-disconfirmation and evaluability literatures.

Practical Implications

Our results suggest a potential downside for platforms' signals of quality. While these signals have been shown to increase demand in well-controlled studies, our results suggest that this effect on demand is likely to be dampened in marketplaces with both ratings and certifications by the competing effect that signals of quality have on ratings. This is because consumers also rely on user-generated ratings when making choices. In our follow-up studies, we find that this dampening can be quite severe, as consumers' choices followed star ratings more strongly than they followed the platform certification.

This does not mean that platforms should simply abandon certifications of quality. Platforms have tools to counteract the possibility for certifications to reduce demand. The most obvious set of tools influence consumer search. For example, Airbnb allows consumers to filter search results by superhost status (as does eBay with top rated seller designations). Even more strongly, Spotify creates playlists out of certified songs and artists, reducing friction for choosing certified songs while increasing it for uncertified ones. Unsurprisingly, certifications lead to large increases in popularity on Spotify (Aguiar and Waldfogel 2018; note that Spotify also does not include ratings). Our results add to the notion that affecting search is a key benefit of certifications, as this allows platforms to avoid consumers directly comparing certified and uncertified alternatives.

Extension to Other Marketplaces

Our investigation has been limited to the effect of superhost status on Airbnb ratings and choices. However, our hypotheses are not specific to this context. For example, we would anticipate other certifications to have similar effects in other contexts, as long as those certifications adjust the comparisons raters make. Return to an example from our introduction: eBay Top Rated Sellers (note that despite the name, this designation is largely not based on user-generated ratings). The decentralized nature of eBay means that customers often wait a long time to receive their products after purchase. Experienced customers will be aware of this, and expect waiting for most purchases. However, people likely expect top rated sellers to ship their products more promptly and efficiently. Thus, waiting for two weeks to receive shipment from a top rated seller will almost certainly lead to a more negative rating than the same wait from a regular seller. Note that this does not necessarily require consumers to have clear *a priori* expectations for shipping time. If the top rated seller designation merely causes consumers to think "How long was this wait, compared to other top rated sellers?", the same detrimental effect of status should arise.

Similar situations should arise in any context wherein an objectively high-quality alternative is over-hyped, or causes consumers to think of high quality alternatives when rating. Doctors who win professional awards may receive low ratings on HealthGrades.com for not working miracles, while unawarded doctors may receive higher ratings because patients do not come to them for miracles. Or objectively superior vehicles may receive lower ratings by missing lofty expectations—despite outperforming the competition.

This becomes a problem when prospective consumers do not understand the factors behind a rating. Travelers may avoid good accommodations, drivers may avoid good cars, and patients may avoid better doctors if they over-rely in user-generated ratings as a point of comparison. In many cases, consumers cannot be expected to understand the factors behind ratings. Awareness would require consumers to put themselves into the mindset of those creating ratings, which is entirely inconsistent with the mindset of "which product should I buy?" Even if consumers can be aware of these factors, our data suggest they are not. This suggests that consumers' default belief is that differences in ratings convey meaningful differences between the alternatives themselves.

Mitigating Effects on Ratings

Compared to the issue we raise, it is relatively more clear how platforms can mitigate other previously identified concerns with user-generated ratings. The prevalence of "fake" reviews creates a degree of uncertainty (Anderson and Simester 2014; Luca and Zervas 2016; Mayzlin, Dover and Chevalier 2014; Stern 2018), but more rigorous standards for posting limit their impact. Issues like small sample size (de Langhe et al. 2016; Powell et al. 2017), self-selection (Bondi 2019; Bondi and Stevens 2019; Li and Hitt 2008), and ulterior motives of raters (Hu, Zhang and Pavlou 2009; Schoenmueller, Netzer and Stahl 2020) can be overcome by encouraging a larger and more representative sample to rate.

Meanwhile, the issue we raise is inherent to ratings' creation. We are not aware of research identifying a reliable, well-founded intervention that would remove the effect of mental context—operationalized here through platform certifications—on ratings. While platforms could attempt to diminish this effect through the information they present at the time of rating (i.e., not showing superhost status when raters evaluate Airbnb listings), this will have limited impact because context affects the perceived experience, not just the rating. Moreover, platforms likely cannot successfully mitigate this by explicitly highlighting that certifications can reduce ratings, as this would require very precise calibration. Nonetheless, we think this may be the most reasonable intervention to begin researching.

People use ratings in part because they are easy to compare across alternatives, but consumers are known to strongly weigh many other forms of information that seem comparable across alternatives (Kivetz and Simonson 2000; Nowlis and Simonson 1997; Slovic and MacPhillamy 1974). Therefore, platforms could work to make objective information more easily understandable, comparable, and accessible for between- product comparisons. We think a good starting point is the comparison feature that many sites provide to shoppers. These tools could cull the attributes presented to be specific to shoppers' needs, more understandable (e.g., screen size would be more useful if presented as scaled diagrams), and accurate (e.g., "Operating System Compatibility" indicates a difference that does not exist). Additionally, platforms could cultivate ratings from experts—who may not be swayed by quality designations—and present them alongside user-generated star ratings, making the cost for consumers to acquire expert information equal to the cost of acquiring user-generated ratings.

Future Research

The current manuscript focuses on consumers' evaluations and interpretations of information online. The results we observe raise important economic issues we hope are addressed in future research. For example, we are unable to estimate the direct economic impact of our results in Study 1 on Airbnb hosts. While the ATTs of –.045 for gaining and .107 for losing superhost status are substantial effects on ratings—representing 8.9% and 21.1% of the standard deviation in Airbnb ratings for these listings—we are unable to estimate the dollar cost of these effects on hosts, as we do not have booking data. Future research could make use of such data to quantify these costs and benefits.

Further, and relatedly, future research could examine the effects of status and ratings on choice independently. We designed Studies 2B and 3 as conservative tests of the effects of ratings and status on choice, which precluded us from quantifying the effect of each independently. Future research could benefit from separating these effects and investigating settings and types of status designations that have stronger and weaker effects on ratings and choice. One specific case in which this comparison could be enlightening is among new offerings from certified providers. For example, because superhost status is a host-level designation on Airbnb, it is possible for brand new listings—with no ratings—to be designated as superhosts. In this case, it is likely that the superhost designation has a positive effect on demand by increasing consumer confidence, in line with Watson, Ghosh, and Trusov's (2018) finding that review counts are more influential than rating levels for new offerings. However, it is also possible that the effect of status on ratings is even stronger in this scenario, as status may be one of the only signals available to consumers, expanding the gap in expectations.

Our investigation also highlights an understudied tension between how ratings are created and used by consumers. Prospective consumers often use ratings to compare specific alternatives, whereas raters do not consider these alternatives when creating their ratings. Instead, raters use internal aspects of products (e.g., expectations) to inform their judgment of what is "good" or "bad" performance. In this manuscript, we find that this fact can lead systematic differences in star ratings to arise absent differences in quality. Moreover, we argue (with H_2) that prospective consumers are insufficiently aware of this possibility, leading to differences in ratings that could have meaningful effects on choice and welfare.

The theoretical implication of this extends beyond signals that platforms provide. Any attribute of a product that affects consumers' expectations, the alternatives they compare the product to, or how they make those comparisons, in turn should affect the ratings for that product. Therefore, we hope that this work provides a beginning substantive thrust for research to consider a broader range of context effects on ratings. We have identified merely one case where the context of consumers' ratings (expectations, which are formed by platform certifications) creates differences in ratings between products. There is no reason to suggest that this is limited to our substantive finding. In particular, future research could assess the role of prior ratings as a source of expectations. While Godes and Silva (2012) find that ratings fall over time, they demonstrate this across all rating levels—both high and low rated products fall over time, because later consumers cannot assess the diagnosticity of prior reviews. Instead, future work could assess an interaction between prior rating level and future ratings.

CONCLUSION

Our results suggest a downside to platform-created certifications. While these signals have been shown to increase demand in prior research, we find that they decrease ratings. Troublingly, prospective consumers under-correct for the influence of these signals on ratings. As a result, platforms' certifications are not as effective as possible. We hope future work will expand on this work in two specific areas. First, research could identify the theoretical underpinnings of this result by discussing the seeming misalignment between ratings' creation and use. Second, broader work could distinguish between the two causes we discuss for the detrimental effect of signals on ratings—to what extent is this a demonstration of expectation-disconfirmation, versus differences in specific alternatives consumers compare experiences to?

REFERENCES

Aguiar, Luis, and Joel Waldfogel (2018). “Platforms, promotion, and product discovery: Evidence from Spotify playlists,” *National Bureau of Economic Research*, (No. w24713).

Airbnb (2024), “Understanding the Superhost Program.” https://www.airbnb.com/help/article/828#:~:text=A%20Superhost%20is%20a%20host,their%20Airbnb%20listing%20and%20profile.

Anderson, Eric T. and Duncan I. Simester (2014), “Reviews Without a Purchase: Low Ratings, Loyal Customers, and Deception,” *Journal of Marketing Research*, 51(3), 249–69.

Askalidis, Georgios, Su Jung Kim, and Edward C. Malthouse (2017), “Understanding and Overcoming Biases in Online Review Systems,” *Decision Support Systems*, 97, 23–30.

Bearden, William O. and Jesse E. Teel (1983), “Selected Determinants of Consumer Satisfaction and Complaint Reports,” *Journal of Marketing Research*, 20(1), 21.

Birnbaum, Michael H. (1999), “How to Show That 9 \Textgreater 221: Collect Judgments in a Between-Subjects Design,” *Psychological Methods*, 4(3), 243–49.

Blanchard, Simon J, Jacob Goldenberg, Koen Pauwels, and David A Schweidel (2022), “Promoting Data Richness in Consumer Research: How to Develop and Evaluate Articles with Multiple Data Sources,” *Journal of Consumer Research*, 49(2), 359–72.

Bondi, Tommaso (2019), “Alone, Together: Product Discovery Through Consumer Ratings,” {{NET Institute Working Paper}}, Rochester, NY.

Bondi, Tommaso and Ryan Stevens (2019), “The Good, The Bad and The Picky: Consumer Heterogeneity and The Reversal of Movie Ratings.”

Callaway, Brantly and Pedro H. C. Sant’Anna (2018), “Did: Treatment Effects with Multiple Periods and Groups,” 2.1.2.

Callaway, Brantly and Pedro HC Sant'Anna (2021), "Difference-in-Differences with Multiple Time Periods," *Journal of econometrics*, 225(2), 200–230.

Chen, Yubo, Qi Wang, and Jinhong Xie (2011), "Online Social Interactions: A Natural Experiment on Word of Mouth Versus Observational Learning," *Journal of Marketing Research*, 48(2), 238–54.

Chintagunta, Pradeep K., Shyam Gopinath, and Sriram Venkataraman (2010), "The Effects of Online User Reviews on Movie Box Office Performance: Accounting for Sequential Rollout and Aggregation Across Local Markets," *Marketing Science*, 29(5), 944–57.

Churchill Jr, Gilbert A and Carol Surprenant (1982), "An Investigation into the Determinants of Customer Satisfaction," *Journal of marketing research*, 19(4), 491–504.

Coulthard, Lisa J Morrison (2004), "A Review and Critique of Research Using SERVQUAL," *International Journal of Market Research*, 46(4), 479–97.

de Langhe, Bart, Philip M. Fernbach, and Donald R. Lichtenstein (2016), "Navigating by the Stars: Investigating the Actual and Perceived Validity of Online User Ratings," *Journal of Consumer Research*, 42(6), 817–33.

Dellarocas, Chrysanthos, Xiaoquan (Michael) Zhang, and Neveen F. Awad (2007), "Exploring the Value of Online Product Reviews in Forecasting Sales: The Case of Motion Pictures," *Journal of Interactive Marketing*, 21(4), 23–45.

Elfenbein, Daniel W, Raymond Fisman, and Brian McManus (2015), "Market Structure, Reputation, and the Value of Quality Certification," *American Economic Journal: Microeconomics*, 7(4), 83–108.

Fleischer, Aliza, Eyal Ert, and Ziv Bar-Nahum (2022). “The Role of Trust Indicators in a Digital Platform: A Differentiated Goods Approach in an Airbnb Market,” *Journal of Travel Research*, 61(5), 1173-1186. https://doi.org/10.1177/00472875211021660

Godes, David, and José C. Silva (2012). “Sequential and temporal dynamics of online opinion,” *Marketing Science*, *31*(3), 448-473.

Grönroos, Christian (1982), “An Applied Service Marketing Theory,” *European journal of marketing*, 16(7), 30–41.

Hsee, Christopher K. (1996), “The Evaluability Hypothesis: An Explanation for Preference Reversals Between Joint and Separate Evaluations of Alternatives,” *Organizational Behavior and Human Decision Processes*, 67(3), 247–57.

Hu, Nan, Jie Zhang, and Paul A. Pavlou (2009), “Overcoming the J-shaped Distribution of Product Reviews,” *Communications of the ACM*, 52(10), 144–47.

Hui, Xiang, Zekun Liu, and Weiqing Zhang (2023). “From high bar to uneven bars: The impact of information granularity in quality certification,” *Management Science*, *69*(10), 6109-6127.

Hui, Xiang, Maryam Saeedi, Zeqian Shen, and Neel Sundaresan (2016), “Reputation and Regulations: Evidence from eBay,” *Management Science*, 62(12), 3604–16.

Kahneman, Daniel (2011), *Thinking, Fast and Slow*, Macmillan.

Kivetz, Ran and Itamar Simonson (2000), “The Effects of Incomplete Information on Consumer Choice,” *Journal of marketing research*, 37(4), 427–48.

Lewis, Gregory (2011), “Asymmetric Information, Adverse Selection and Online Disclosure: The Case of eBay Motors,” *American Economic Review*, 101(4), 1535–46.

Lewis, Robert C. and Bernard H. Booms (1983), "The Marketing Aspects of Service Quality," in *Emerging Perspectives on Services Marketing*, L. Berry, G. Shostack, and G. Upah, eds., Chicago: American Marketing, 99–107.

Li, Shibo, Kannan Srinivasan, and Baohong Sun (2009), "Internet Auction Features as Quality Signals," *Journal of Marketing*, 73(1), 75–92.

Li, Xingyi, Yiting Deng, Puneet Manchanda, and Bert De Reyck (2022), "Can Lower Expert Opinions Lead to Better Consumer Ratings?: The Case of Michelin Stars."

Li, Xinxin and Lorin M. Hitt (2008), "Self-Selection and Information Role of Online Product Reviews," *Information Systems Research*, 19(4), 456–74.

Luca, Michael and Oren Reshef (2021), "The Effect of Price on Firm Reputation," *Management Science*, 67(7), 4408–19.

Luca, Michael and Georgios Zervas (2016), "Fake It Till You Make It: Reputation, Competition, and Yelp Review Fraud," *Management Science*, 62(12), 3412–27.

Lynch, John G., Howard Marmorstein, and Michael F. Weigold (1988), "Choices from Sets Including Remembered Brands: Use of Recalled Attributes and Prior Overall Evaluations," *Journal of Consumer Research*, 15, 169–84.

Mayzlin, Dina, Yaniv Dover, and Judith Chevalier (2014), "Promotional Reviews: An Empirical Investigation of Online Review Manipulation," *American Economic Review*, 104(8), 2421–55.

Mishra, Rajan, Guofang Huang, and Manohar Kalwani (2023), "The Value of Reputation Badges for Sellers in the Age of Ratings and Review: An Empirical Study of Airbnb's Superhost Program," *SSRN Electronic Journal*.

Nowlis, Stephen M. and Itamar Simonson (1997), “Attribute–Task Compatibility as a Determinant of Consumer Preference Reversals,” *Journal of marketing research*, 34(2), 205–18.

Oliver, Richard L. (1977), “Effect of Expectation and Disconfirmation on Postexposure Product Evaluations: An Alternative Interpretation.” *Journal of Applied Psychology*, 62(4), 480–86.

——— (1980), “A Cognitive Model of the Antecedents and Consequences of Satisfaction Decisions,” *Journal of Marketing Research*, 17(4), 460–69.

Oliver, Richard L and Wayne S DeSarbo (1988), “Response Determinants in Satisfaction Judgments,” *Journal of consumer research*, 14(4), 495–507.

Parasuraman, ABLL, Valarie A Zeithaml, and L Berry (1988), “SERVQUAL: A Multiple-Item Scale for Measuring Consumer Perceptions of Service Quality,” *1988*, 64(1), 12–40.

Parasuraman, A., Valarie A. Zeithaml, and Leonard L. Berry (1985), “A Conceptual Model of Service Quality and Its Implications for Future Research,” *Journal of Marketing*, 49(4), 41–50.

Powell, Derek, Jingqi Yu, Melissa DeWolf, and Keith J. Holyoak (2017), “The Love of Large Numbers: A Popularity Bias in Consumer Choice,” *Psychological Science*, 28(10), 1432–42.

R Core Team (2024), *R: A Language and Environment for Statistical Computing*, Manual, Vienna, Austria: R Foundation for Statistical Computing.

Rietveld, Joost, Robert Seamans, and Katia Meggiorin (2021), “Market Orchestrators: The Effects of Certification on Platforms and Their Complementors,” *Strategy Science*, 6(3), 244–64.

Rossi, Michelangelo (2021), “Quality Disclosures and Disappointment: Evidence from the Academy Awards,” in *Proceedings of the 22nd ACM Conference on Economics and Computation*, Budapest Hungary: ACM, 790–91.

Rust, Roland T. and Richard L. Oliver (1994), “Service Quality: Insights and Managerial Implications from the Frontier,” *Service Quality: New Directions in Theory and Practice*, 1–20.

Schoenmueller, Verena, Oded Netzer, and Florian Stahl (2020), “The Polarity of Online Reviews: Prevalence, Drivers and Implications,” *Journal of Marketing Research*, 57(5), 853–77.

Simonson, Itamar (2016), “Imperfect Progress: An Objective Quality Assessment of the Role of User Reviews in Consumer Decision Making, A Commentary on de Langhe, Fernbach, and Lichtenstein,” *Journal of Consumer Research*, 42, 840–45.

Slovic, Paul (1972), “From Shakespeare to Simon: Speculations–and Some Evidence–about Man’s Ability to Process Information.”

Slovic, Paul and Douglas MacPhillamy (1974), “Dimensional Commensurability and Cue Utilization in Comparative Judgment,” *Organizational Behavior and Human Performance*, 11(2), 172–94.

Stern, Joanna (2018), “Is It Really Five Stars? How to Spot Fake Amazon Reviews,” *Wall Street Journal*.

Watson, Jared, Anastasiya P. Ghosh, and Michael Trusov (2018), “Swayed by the numbers: the consequences of displaying product review attributes,” *Journal of Marketing*, 82(6), 109-131.

Woodruff, Robert B, Ernest R Cadotte, and Roger L Jenkins (1983), “Modeling Consumer Satisfaction Processes Using Experience-Based Norms,” *Journal of marketing research*, 20(3), 296–304.

Yao, Bin, Richard T. R. Qiu, Daisy X. F. Fan, Anyu Liu, and Dimitrios Buhalis (2019), “Standing Out from the Crowd – an Exploration of Signal Attributes of Airbnb Listings,” *International Journal of Contemporary Hospitality Management*, 31(12), 4520–42.

WEB APPENDICES

WEB APPENDIX A: GOODREADS.COM ANALYSIS

We obtained Goodreads ratings data in March 2022 from Kaggle.com.[1] We restricted our sample to books published between 1917 and 2022 with at least 12 ratings (the median of the total set), and combined ratings for books that had multiple versions. Pulitzer prize winners were designated by scraping prize information from Wikipedia at the same time.

We identified all books in this set that had won a Pulitzer Prize since 1917, about which we make two assumptions: (i) The prize-winning books are of high quality and (ii) readers will consume these books with high expectations. If star ratings only reflect the quality, we should expect the winners to have higher than average ratings.[2] Instead, we find that ratings of winners and non-winners are similar ($M_{\text{PrizeWinners}}$ = 4.00/5 vs. M_{AllBooks} = 3.89/5) and ranks in the 59th percentile of all books in its publication year. It appears then that Pulitzer Prize winning books are rated within the context of being "the best book of the year," while others are not rated against such a high bar. As a result, many books that are (likely) objectively worse are rated higher than prize winners. Pushing our logic to an extreme, a consumer who only used Goodreads.com user-ratings to select books would have to read 59,037 other books before reading Grapes of Wrath (Pulitzer Prize in 1940, rated 3.97/5 on Goodreads.com), which was cited as a "great work" in the decision to award John Steinbeck a Nobel Prize for Literature (Österling 1962) and has been featured on numerous lists of "best novels" (Grossman and Lacayo 2010; BBC 2003).

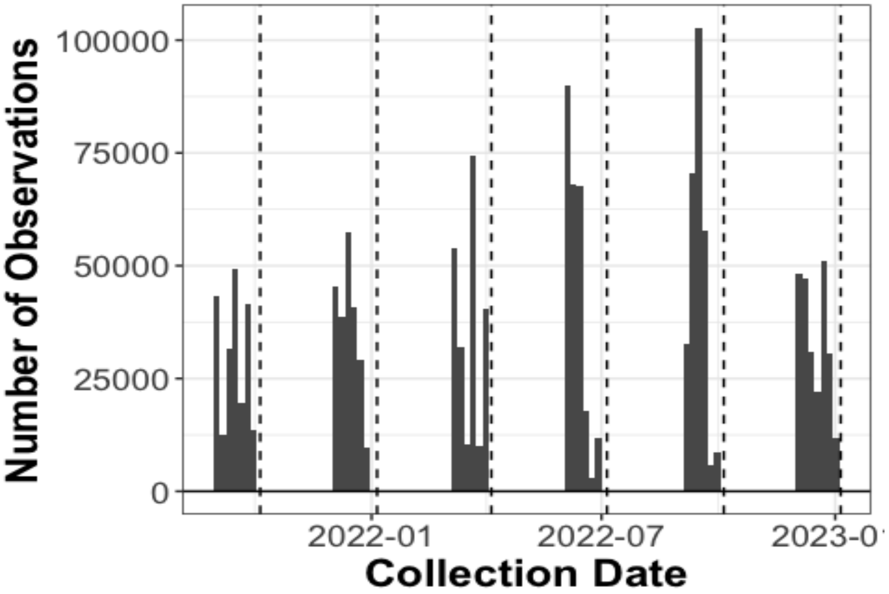
[1] https://www.kaggle.com/datasets/bahramjannesarr/goodreadsbook- datasets-10m/activity

[2] One obvious critique–and the reason we do not claim strong causal inference from these data–is that awards also influence choice. As demonstrated by Bondi (2019), it is likely that consumers who are less inclined to read a given book (due to a preference match) are more likely to read it after being awarded. These consumers will get less utility out of the book regardless of expectations.

WEB APPENDIX B: REWARDING OF SUPERHOST STATUS, COMPARED TO INSIDEAIRBNB OBSERVATIONS

Our listing snapshots come from six quarters of data collected by InsideAirbnb. These collection dates are noted by InsideAirbnb, and come immediately before the rewarding of superhost status, as evidenced by Figure B1. Airbnb awards Superhost status on January 1, April 1, July 1, and October 1 of every year.

FIGURE B1
DISTRIBUTION OF INSIDEAIRBNB OBSERVATION DATES



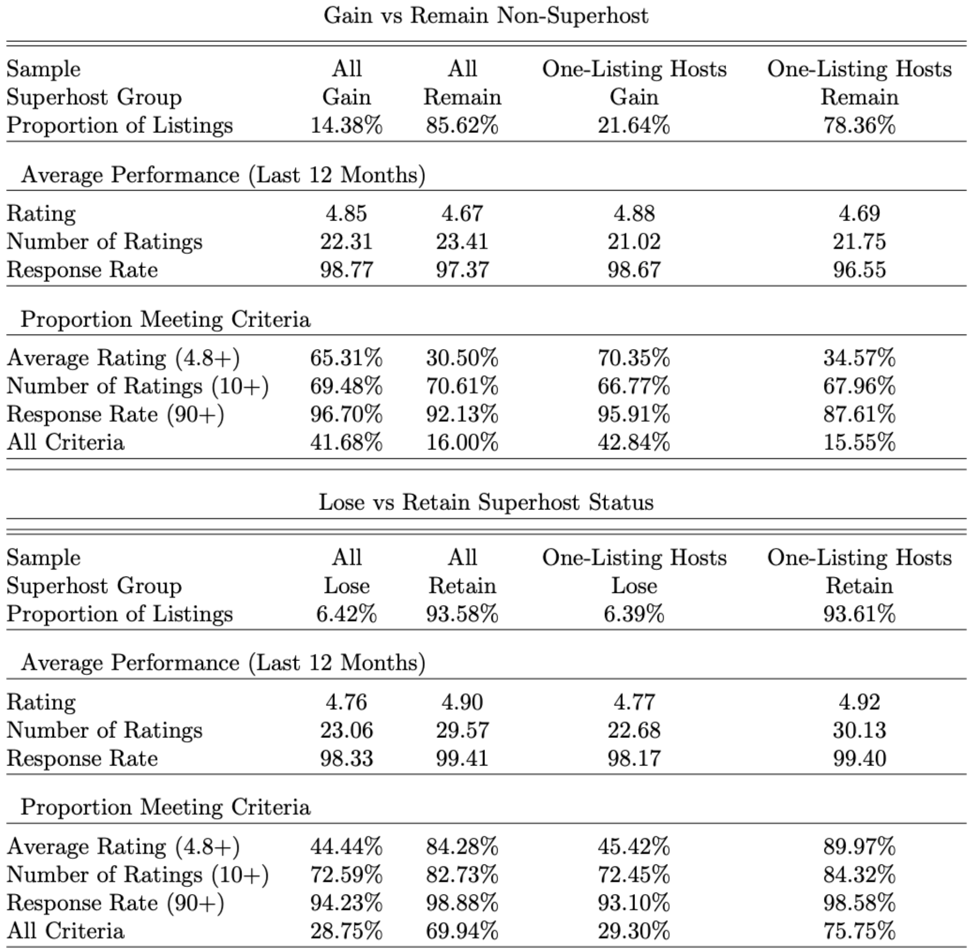
NOTE.—Vertical lines indicate Airbnb Superhost change dates.

Because these collections take place immediately before status changes are announced, we consider these InsideAirbnb observations to be summaries of the prior quarter. We assume that all information in these snapshots is relevant to that quarter. This means that any ratings provided during that quarter should also have been provided under the same superhost status as observed by InsideAirbnb.

WEB APPENDIX C: CHANGES IN SUPERHOST STATUS

In a given quarter, 14.4% of non-superhosts gain superhost status. Gaining status is correlated with Airbnb's posted criteria, though imperfectly so. Of the listings who gain status in a given quarter, only 65.3% achieved an average rating of 4.8 or higher in the prior year. Only 69.5% received 10 or more ratings in the prior year (which we use as a conservative estimate for the 10 bookings needed), while 96.7% met the response rate threshold of 90%. Together, 41.7% of listings who gain superhost status met all three criteria in the prior year.

TABLE C1
PERFORMANCE AGAINST SUPERHOST CRITERIA



Gain vs Remain Non-Superhost

Sample	All	All	One-Listing Hosts	One-Listing Hosts
Superhost Group	Gain	Remain	Gain	Remain
Proportion of Listings	14.38%	85.62%	21.64%	78.36%
Average Performance (Last 12 Months)				
Rating	4.85	4.67	4.88	4.69
Number of Ratings	22.31	23.41	21.02	21.75
Response Rate	98.77	97.37	98.67	96.55
Proportion Meeting Criteria				
Average Rating (4.8+)	65.31%	30.50%	70.35%	34.57%
Number of Ratings (10+)	69.48%	70.61%	66.77%	67.96%
Response Rate (90+)	96.70%	92.13%	95.91%	87.61%
All Criteria	41.68%	16.00%	42.84%	15.55%

Lose vs Retain Superhost Status

Sample	All	All	One-Listing Hosts	One-Listing Hosts
Superhost Group	Lose	Retain	Lose	Retain
Proportion of Listings	6.42%	93.58%	6.39%	93.61%
Average Performance (Last 12 Months)				
Rating	4.76	4.90	4.77	4.92
Number of Ratings	23.06	29.57	22.68	30.13
Response Rate	98.33	99.41	98.17	99.40
Proportion Meeting Criteria				
Average Rating (4.8+)	44.44%	84.28%	45.42%	89.97%
Number of Ratings (10+)	72.59%	82.73%	72.45%	84.32%
Response Rate (90+)	94.23%	98.88%	93.10%	98.58%
All Criteria	28.75%	69.94%	29.30%	75.75%

This summary is presented in Table C1, alongside listings that remain non-superhosts in consecutive quarters. Table C1 also demonstrates consistency among hosts with only one listing, which we consider because superhost is decided at the host level. While there are clear differences between groups in performance, a large proportion of listings gain superhost status without meeting all criteria. We see similar lack of clarity when describing the listings that lose superhost status in the bottom panel of Table C1: Results are largely consistent with what we might expect from the listed criteria, but are far from clear. We would expect more listings that meet all criteria to gain superhost status, and more listings that do not meet all criteria to lose superhost status. Thus, unobserved variables must be impacting changes in superhost status.

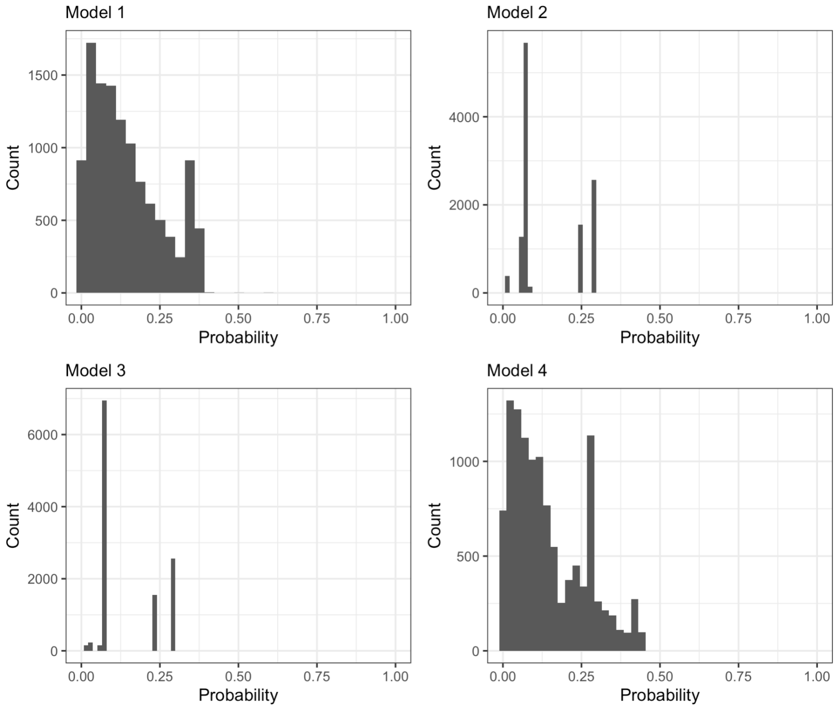
Predicting Status Changes

A potential identification strategy that handles the fact that listings influence their own treatment would be to compare listings that "barely" changed to those who "nearly" changed. In cases where treatment is determined by a combination of factors (such as ours), this requires modeling the likelihood of being treated (i.e., the likelihood of changing status in a given quarter), and only comparing ratings for listings that were highly likely to change status in our difference-in-differences. Unfortunately, this strategy is not possible as a direct result of the lack of clarity in superhost status determination on Airbnb.

We attempted to model changes in superhost status, using observable criteria. We compared five models of gaining and losing status separately. Each model was trained on 70% of the sample of eligible listing-quarter observations (28,573 observations of non-superhosts to predict gaining status, and 55,904 observations of superhosts to predict losing status). We then tested each model by comparing its predictions for a hold out set of the other 30% of observations (12,246 observations of non-superhosts to predict gaining status, and 23,959 observations of superhosts to predict losing status).

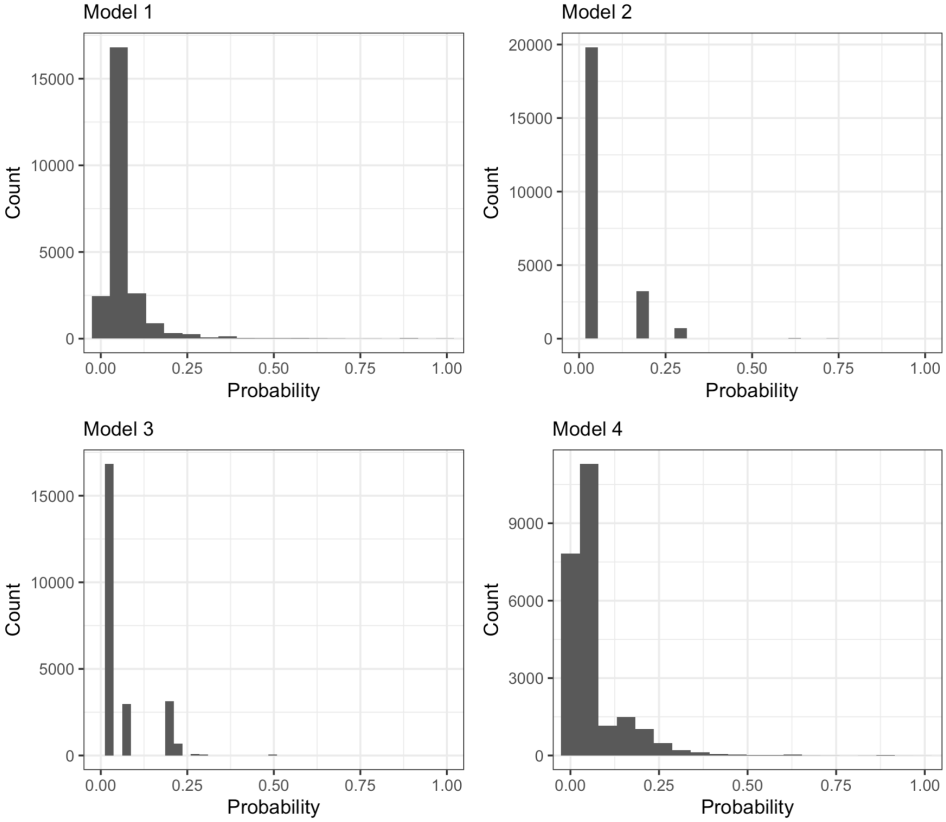
The first four models are logistic regressions, predicting change (either gain or loss) as a function of annual average ratings, average response rate, and number of ratings. Model 1 considers these as continuous predictors, and only includes main effects. Model 2 implements these variables as dummy codes, indicating whether or not the superhost threshold was met. Model 2 only includes main effects. Model 3 replicates Model 2, but includes all interactions. Model 4 includes all predictors in Model 3, but also includes main effects for the continuous versions of each variable. Finally, Model 5 simply makes a random prediction according to the overall frequency of gaining (14.4%) and losing (6.4%) status.

FIGURE C1
PREDICTED PROBABILITY OF GAINING SUPERHOST STATUS IN TEST SET



Results—Gaining Status. As observed in Figure C1, no logistic regression model ever predicts above 50% that a listing in the testing set will become a superhost. While this alone illustrates the lack of clarity in superhost designation, we decided to classify a listing as predicted to be a superhost if the model estimate was above the mean incidence of gaining superhost status (14.4%). Each model predicted those who remain non-superhosts accurately (Model 1: 94.54%; Model 2: 92.60%; Model 3: 92.60%; Model 4: 94.09%), but only slightly better than random guessing (85.71%). However, no model predicted those who gained superhosts accurately (Model 1: 27.86%; Model 2: 28.03%; Model 3: 28.03%; Model 4: 28.22%). While each was better than random guessing (12.86%), we are clearly not able to use Airbnb's posted criteria and posted data to predict gaining status.

FIGURE C2
PREDICTED PROBABILITY OF LOSING SUPERHOST STATUS IN TEST SET



Results—Losing Status. As observed in Figure C2, no logistic regression model predicts above 50% that a listing in the testing set will lose superhost status. We classify a listing as predicted to lose status if the model estimate was above the mean incidence of losing superhost status (6.4%). The four models predicted remaining superhosts quite accurately (Model 1:

97.02%; Model 2: 96.63%; Model 3: 97.17%; Model 4: 96.89%), but so did random guessing (93.71%). However, no model predicted those who gained superhosts accurately (Model 1: 15.95%; Model 2: 21.32%; Model 3: 14.92%; Model 4: 18.65%). While better than random guessing (8.26%), this means that the best model only correctly predicted 1/5 of superhost status losses. Thus, we are not able to use Airbnb's posted criteria and data to predict losing status.

Predicting Status Changes Within Narrow Window

Due to our inability to accurately model superhost designations across the entire data, we attempted to do so in a narrow window where ratings should be the most impactful determinant of status. We did so by considering instances where listings met all other criteria, and were near the ratings cut-off.

There were 5,607 instances where a non-superhost listing achieved at least 10 ratings in the prior year, at least a 90% response rate, and average annual ratings between 4.75 and 4.85 (inclusive)—a narrow window where being above and below 4.8 could be considered "as good as random." In this narrow set, 1,433 listings became superhosts in the next quarter, and 4,174 did not. However, the proportion of those whose ratings were at or above 4.8 was extremely close in each (53.45% for those who gained, 49.95% for those who did not). Thus, it does appear that small variations in ratings cause changes in status to a great extent.

Discussion

Because we are not able to accurately model changes in status, we cannot reliably employ an identification strategy that relies on estimated changes. Therefore, we are required to focus on the remaining two within-listing strategies. While these strategies may not completely rule out all concerns, they provide the best causal claim we can make in this context.

WEB APPENDIX D: VRBO MATCHING PROCESS

To complete this task, we scraped Vrbo by searching for the cities for whom we have Airbnb data. This yielded 35,978 listings. We then developed a simple algorithm to match Airbnb listings to themselves on Vrbo. This was challenging, as there is no unique key between the data sets. Both platforms mask listings' exact locations, and there is no requirement for listings' names, descriptions, or host names to be the same across platforms. Many of the amenities listed are also not shared across platforms, even among obvious matches.

Exacerbating these difficulties is the fact that including mismatches in our final data would mire the internal validity of our result. We intend to find an interaction between platform (Airbnb vs. Vrbo) and superhost status, such that Airbnb superhost status has a negative effect on Airbnb ratings, but no effect on Vrbo ratings. This would be expected if our Vrbo matches were poor, as we would not expect any element of an unrelated Airbnb listing to affect Vrbo ratings. This motivation to have only good matches led us to match quite strictly. Specifically, we began with all Airbnb listings for whom we observe more than one quarter of InsideAirbnb data. Then, we found a list of potential Vrbo matches from our set of Vrbo listings. We called any listing a potential match if their listed number of guests accommodated was within 1 on each platform, and their euclidean distance was within .07 of eachother.

This process yielded 13,496,289 unique potential matches, with 190,481 Airbnb listings and 30,411 Vrbo listings having at least one potential match.From here, we call an Airbnb-Vrbo pair a match if they meet at least one of the following seven criteria:[3]

1. Having the exact same listing name: 137
2. Having the exact same host name: 3,170

[3] At any level, if multiple pairs met these criteria, we selected the closest one geographically.

3. Having the exact same description: 65
4. Having the exact same first 500 characters of their description: 104
5. Having a listing name that is a subset of the other platform listing name and a host name that is a subset of the other platform host name: 2,080
6. Having a listing name that is a subset of the other platform listing name: 2,865
7. Having a host name that is a subset of the other platform host name: 5,031

This strict matching left us with 13,452 matches we feel confident in. These matches correspond to 104,017 observations for Airbnb and 23,283 for Vrbo.

WEB APPENDIX E: REGRESSION TO THE MEAN IN STUDY 1

Ratings are prone to random variation—particularly due to their high average and small samples within listings in a quarter—and ratings impact superhost status. This variation can be caused by either heterogeneity in consumer tastes and how they rate properties, or due fluctuations in the quality of the properties over time. Therefore, it may seem reasonable that some listings gain status due to a period of good luck, or lose status after a period of bad luck. If so, ratings in the next period may drop for superhosts, or rise for non-superhosts, not due to a difference in expectations, but due to a difference in luck. Certain facts of our data speak against this concern, albeit indirectly. In this web appendix, we discuss three such facts. However, we note that Web Appendix C, which demonstrates a weak relationship between slight variation in ratings and status changes, also weakens this concern.

Parallel Pre-Trends Between Airbnb and Vrbo

The parallel pre-trends between Airbnb and Vrbo ratings are inconsistent with a regression-to-the-mean explanation. A regression-to-the-mean account supposes that superhost status is acquired after a period of anomalously high ratings (i.e., they get lucky then revert back to normal). But, if these higher ratings are indeed an stochastic anomaly—in other words, if they are driven by chance factors—it is not clear why we see parallel pre-trends in the Vrbo data. A pure regression-to-the-mean account would predict divergence in pre-trends as well: We should see a chance increase in ratings in the Airbnb data pre-superhost status (and this chance increase is responsible for the acquisition of superhost status) and then regression-to-the-mean after this better-than-average period. But there is no reason that the same random pattern should manifest in the Vrbo data as well. In this Web Appendix, we present a series of further analyses to address this concern in different ways.

Consistency of Figure 1 With Stable Pre-Trends on Airbnb

For example, we see similar model-free patterns in ratings to Figure 1 in-text when we remove listings with high variation in ratings prior to changing status. Specifically, Figure E1 and Figure E2 replicate Figure 1 from the main text, removing the listings with the highest 25% and 50% of pre-treatment standard deviation in monthly ratings.

FIGURE E1
MONTHLY AVERAGE RATINGS FOR AIRBNB LISTINGS, REMOVING HIGHEST 25TH PERCENTILE IN PRE-CHANGE SD

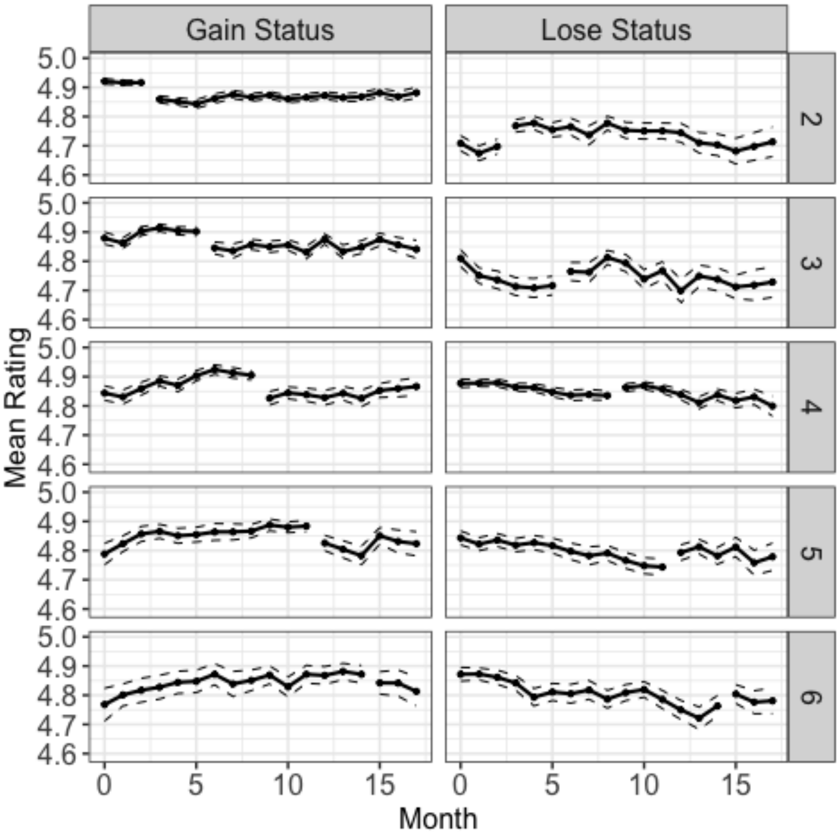
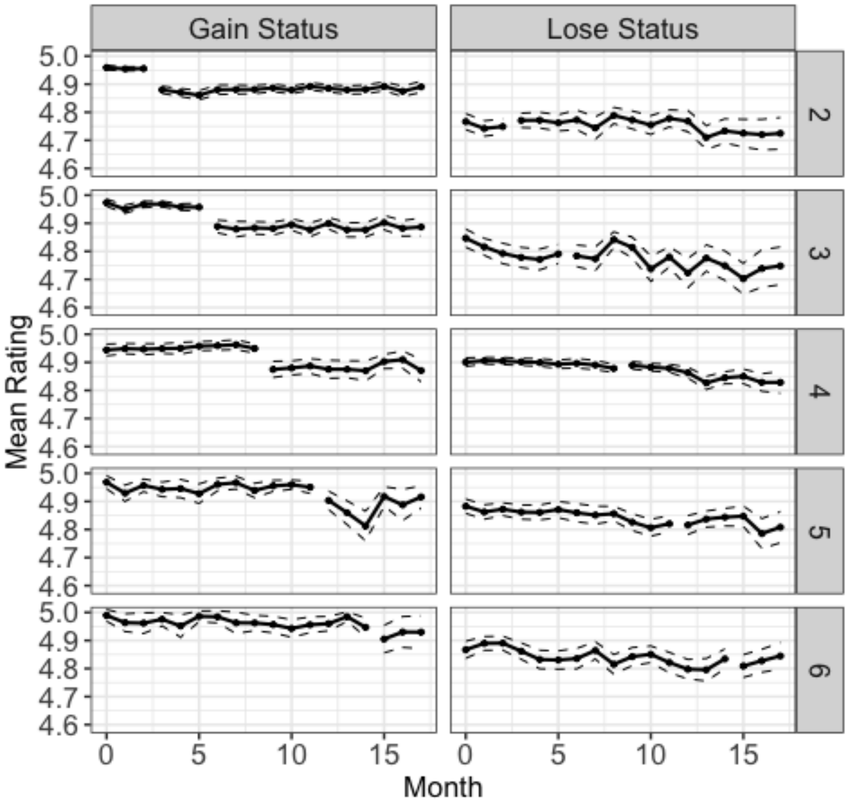


FIGURE E2

MONTHLY AVERAGE RATINGS FOR AIRBNB LISTINGS, REMOVING HIGHEST 50^{TH} PERCENTILE IN PRE-CHANGE SD

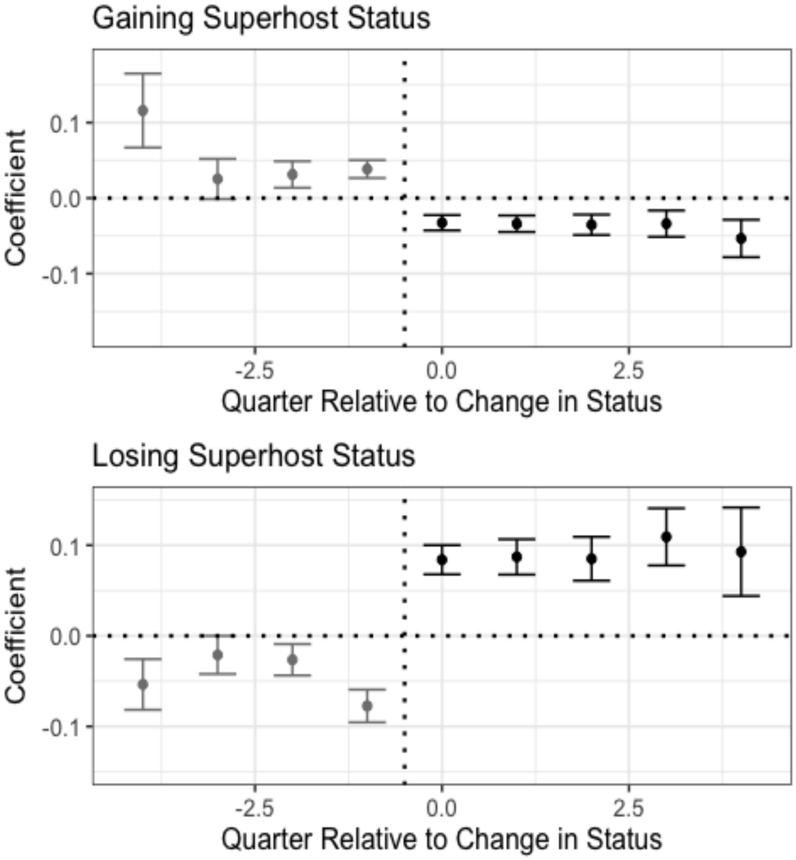


Study 1 Results, Removing Listings Who Both Gain and Lose

The regression to the mean argument suggests that some superhosts are not actually of superhost quality (and vice versa for non-superhosts). As a result, these hosts should be more likely to revert status in the future than other superhosts. This logic suggests that our inclusion of listings who change status more than once would lead to a larger effect of superhost status on ratings, because this includes more listings that are most affected by regression to the mean. This is not to say that all listings affected by regression to the mean will revert status, only that those who revert are more likely to be those that were affected by regression to the mean. To test this, we replicated the results of the first and third identification strategies in Study 1, removing those who both gain and lose status.

Between-Listing Difference in Differences. Among only the listings who change superhost status no more than once, we see similar pre-trends as in text (Figure E3).

FIGURE E3
CALLAWAY AND SANT'ANNA (2021) AVERAGE TREATMENT EFFECTS OF CHANGING SUPERHOST STATUS ACROSS TIME PERIODS



We achieve slightly better results by controlling for hosts' number of listings, observed response rate in the quarter, number of ratings in the quarter, and the listing's price. These results are shown in Figure E4, where we see slightly more parallel pre-trends–evidenced by the coefficients being closer to zero prior to treatment.

FIGURE E4

CALLAWAY AND SANT'ANNA (2021) AVERAGE TREATMENT EFFECTS OF CHANGING SUPERHOST STATUS ACROSS TIME PERIODS, INCLUDING CONTROLS

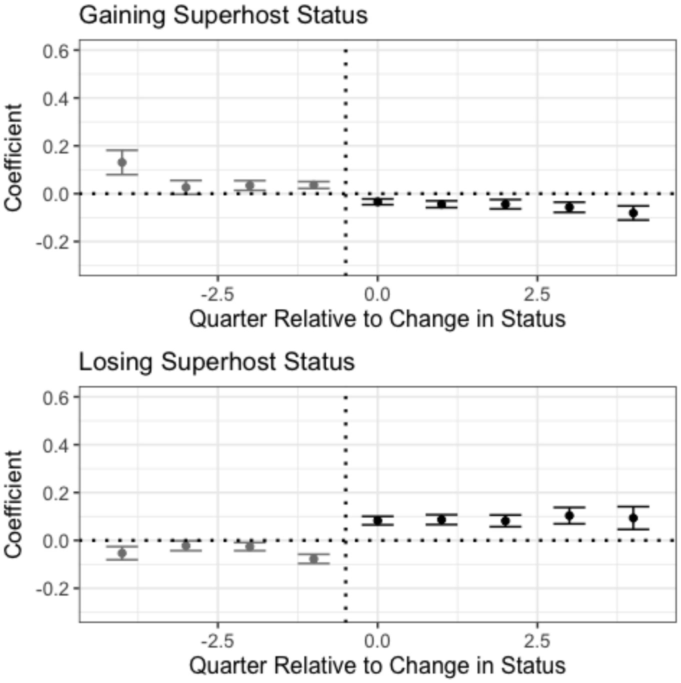


TABLE E1

RESULTS OF DIFFERENCE-IN-DIFFERENCES FOR GAINING AND LOSING STATUS

Model	1	2	3	4
Treated Group	Gain	Gain	Lose	Lose
Control Group	Never	Never	Always	Always
ATT	−.038***	−.052***	.092***	.090***
SE	(.004)	(.006)	(.007)	(.007)
Controls		✓		✓
Listing FEs	✓	✓	✓	✓
Quarter FEs	✓	✓	✓	✓
Observations	437,095	398,435	984,227	983,929
Mean DV	4.72	4.72	4.88	4.88

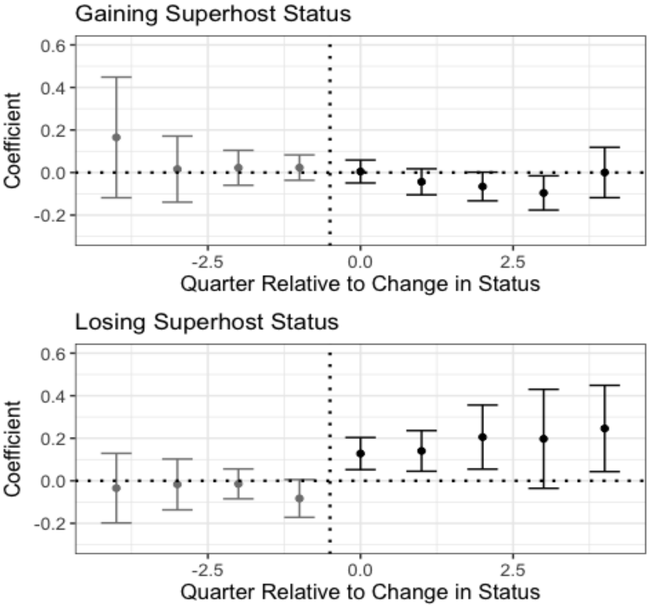
Results support our H_1, and are presented in Table E1 both without controls (Models 1, 3) and including controls (Models 2, 4). Listings who gain superhost status see a more substantial decrease in ratings after gaining relative to those who are never superhosts (ATT = –.037, SE =

.004; Model 1). This remains true after controlling for price, the number of reviews received in that quarter, hosts' response rate, and hosts' number of listings (Model 2), which we use to proxy for quality.

Models 3 and 4 replicate Models 1 and 2 for listings who lose superhost status, comparing them to listings who are always superhosts. Those who lose status see a more substantial increase in ratings after losing (ATT = .09, SE = .006). This is also consistent after controlling for observable attributes of quality (Model 4).

Airbnb-VRBO Difference-in-Differences. In these models, our treated units are the subset of listings whose Airbnb superhost status changes and can be matched to itself on Vrbo. Our control units are those matched units on Vrbo (Figure E5).

FIGURE E5
AVERAGE TREATMENT EFFECTS OF CHANGING SUPERHOST STATUS ACROSS TIME PERIODS



Results with controls are shown in Figure E6, where we see slightly more parallel pre-trends—evidenced by the coefficients being closer to zero prior to treatment.

FIGURE E6
AVERAGE TREATMENT EFFECTS OF CHANGING SUPERHOST STATUS ACROSS TIME PERIODS, INCLUDING CONTROLS

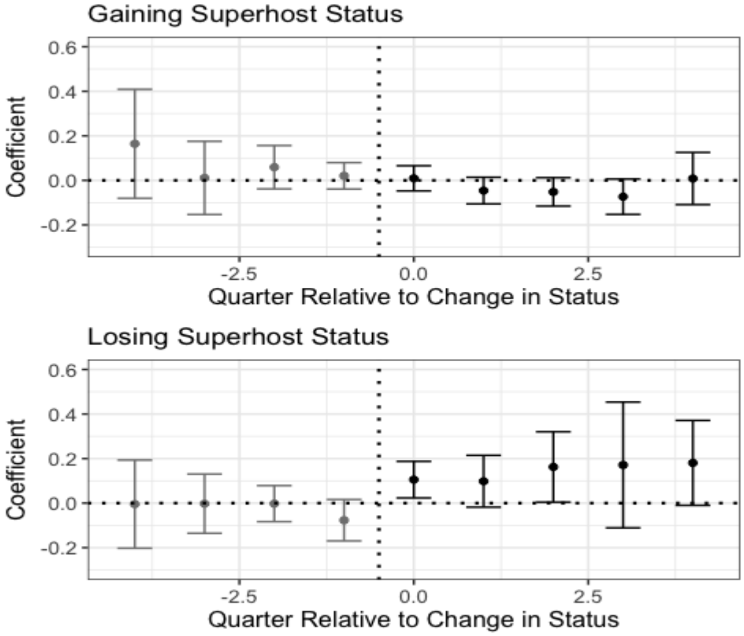


TABLE E2
RESULTS OF DIFFERENCE-IN-DIFFERENCES BETWEEN AIRBNB AND VRBO RATINGS FOR GAINING AND LOSING SUPERHOST STATUS

Model	1	2	3	4
Treated Group	Gain	Gain	Lose	Lose
Control Group	Vrbo	Vrbo	Vrbo	Vrbo
ATT	−.040***	−.031***	.183***	.144***
SE	(.021)	(.021)	(.040)	(.042)
Controls		✓		✓
Listing FEs	✓	✓	✓	✓
Quarter FEs	✓	✓	✓	✓
Observations	14,793	14,780	10,209	10,197
Mean DV	4.85	4.85	4.80	4.80

Results support our H_1. Specifically, listings who gain superhost status see a more substantial decrease in ratings on Airbnb after gaining than they do on Vrbo (Model 1; ATT = –.04, SE = .021). And listings who lose superhost status see a more substantial increase in ratings on Airbnb after losing relative to themselves on Vrbo (Model 3; ATT = .183, SE = .04). We replicate these models with controls for observable listing attributes, which do not affect the results (Models 2 and 4).

Discussion

While no analysis in this section perfectly addresses regression-to-the-mean, they combine to sufficiently narrow the opportunity for this concern. Specifically, Web Appendix C already demonstrates that ratings have a limited impact on superhost status, meaning that slight variation in ratings should not cause large changes in status. This is consistent with the results demonstrated in this appendix. When we limit our data to attempt to exclude listings that are most likely to be prone to regression-to-the-mean, we find no difference in pre-trends (Figures E1, E2) to our main results, and no differences in the effects we observe in either difference-in-differences. Moreover, we note that important inconsistencies with our data to the regression-to-the-mean argument, most pertinently that the Airbnb-Vrbo parallel trends would not be expected if regression-to-the-mean were at play.

WEB APPENDIX F: ROBUSTNESS OF WITHIN-LISTING AIRBNB ANALYSIS

Replicating Within-Listing Models of Airbnb Ratings

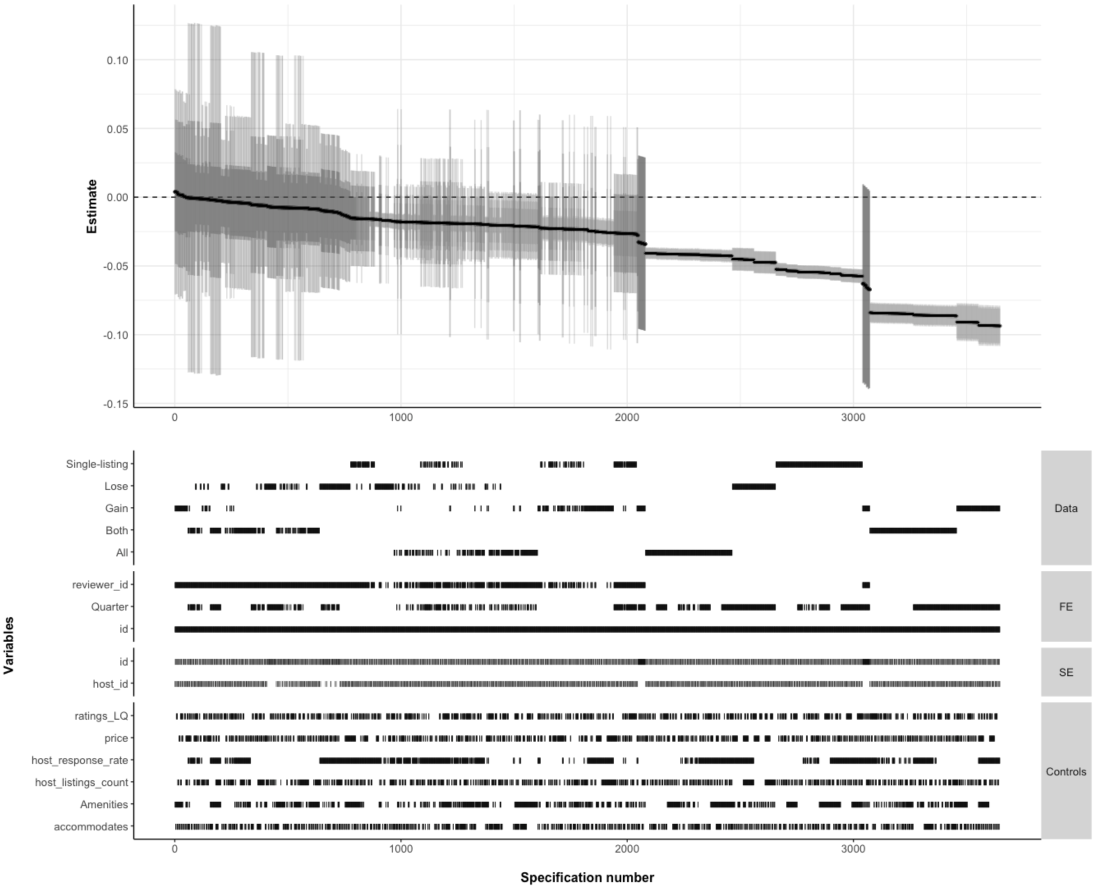
We are limited in the breadth of models we are able to communicate with a table. Therefore, we present a specification curve analysis (Simonsohn, Simmons, and Nelson 2020), in which we investigate the coefficient of superhost status on ratings from 3,840 variants of the focal model. Each variant specification is a unique combination of choices of 1) data, 2) control variables, 3) fixed-effects, and 4) standard error clustering we consider to be reasonable variations of our main model.

Each specification always includes a listing fixed-effect, as results from a model without would be influenced by between-listing differences in quality. We vary whether we include time (quarter) fixed effects to control for overall time trends and reviewer fixed-effects to control for selection. For controls, we vary all combinations of eight possible variables: the number of listings a host has in a given quarter, the number of ratings at a listing in the last quarter (to proxy for demand), the host's response rate to potential guests, the number of people a listing accommodates, the listing's price (winsorized), and the number of amenities listed. Note that each of these attributes vary across time (with the exception of accommodation), allowing us to address time-varying quality for the first time in this identification. Finally, we vary whether we clustered standard errors on hosts, listings, or neither.

These combinations of controls, fixed effects and standard error clusterings corresponds to 768 unique models. Each of these models is then run on one of five data sets: The entire data, the subset of data from hosts with only a single listing (N_{obs} = 266,741), the subset of all listings who only gain superhost status (N_{obs} = 171,436), those who only lose status (N_{obs} = 132,747), and those who do both (N_{obs} = 137,009).

Unfortunately, the inclusion of reviewer fixed-effects leads to nearly perfect fit in many models that include controls. Thus, we remove 192 models whose confidence intervals are in the widest 95th percentile of all. This leaves us with 3,648 total models, of which 1,728 include a reviewer fixed-effect.

FIGURE F1
RESULTS OF STUDY 1 SPECIFICATION CURVE



As shown in Figure F1, most specifications are consistent with our H$_1$, showing a negative estimated effect of superhost status (N_{Models} = 3,600; 98.68% of all models). The median coefficient estimate is –.024, with a median 95% confidence interval between –.050 and –.013. Further, 2,253 (61.2%) of all models had superhost coefficient estimates significantly below zero. This includes 100% of models without reviewer fixed-effects. Meanwhile, 48 models (1.32%) had positive coefficients, with none having statistically significantly positive estimates.

This heterogeneity seems caused largely by models with a reviewer fixed-effect. These models have less negative effects on average, which may be due to nearly perfect fit. Specifically, all models with positive estimated effects of superhost status include a reviewer fixed-effect, which removes much of the variation in ratings. In fact, these models have an average R-squared of .941, suggesting unreliable estimates.

Among models without a reviewer fixed-effect, the effect of superhost status is rather consistent. It is most positive in the subsample of listings who only lose superhost status, and significantly more negative in models that control for hosts' response rate (β = –.004, t(1,529) = -2.482, p = .013).

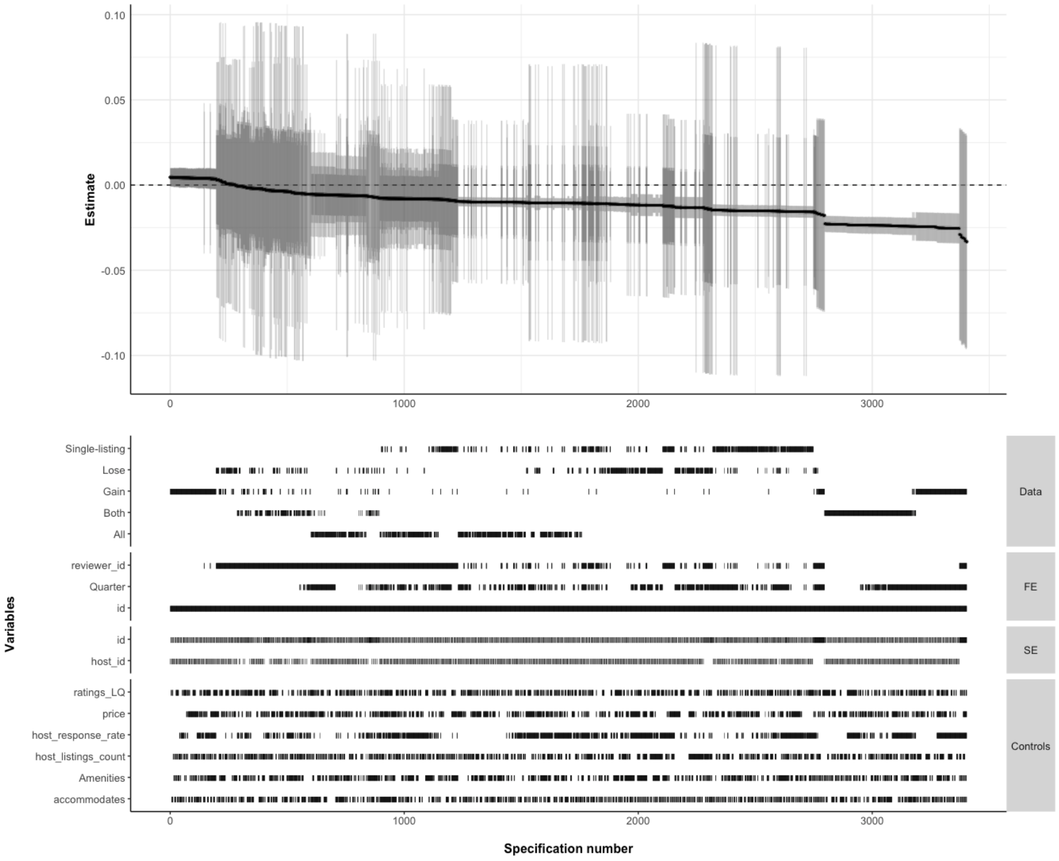
Replicating Within-Listing Models of Airbnb Text Review Sentiment

We also present a specification curve analysis in which we investigate the coefficient of superhost status on *text review sentiment* from 3,840 variants of the focal model. Each specification is a product of the same choices made for the ratings specification curve.

Most specifications are consistent with our H_1, showing a negative estimated effect of superhost status (N_{Models} = 3,127; 91.9% of all models). The median coefficient estimate is –.011, with a median 95% confidence interval between –.028 and –.005. 1,728 (50.8%) of all models had superhost coefficient estimates significantly below zero. This includes 90.0% of models without reviewer fixed-effects. Meanwhile, 277 models (8.1%) had positive coefficients, with none having statistically significantly positive estimates.

FIGURE F2

RESULTS OF STUDY 1 SPECIFICATION CURVE WITH SENTIMENT AS DV



Differences in the Effect of Superhost Status Between Listings

Because our within-listing analysis of Airbnb ratings (Equation 5) includes the most observations and has the most flexible functional form of our three identifications, we can use it to investigate differences in the effect of superhost status between listings in this model. Specifically, we included interactions of superhost status and listing attributes one-at-a-time, testing models of the following equation:

$$Rating_{iq} = \alpha_1 Superhost_{iq} + \alpha_2 Attribute_{iq} + \alpha_3 Superhost \times Attribute_{iq} + \beta X_{iq} + \varepsilon_{iq} \qquad (8)$$

Where $Attribute$ refers to the attribute of a listing considered in each model. Specifically, we repeated this model for each attribute included in the specification curve analysis. Each model included fixed-effects for listing and quarter (but not reviewer due to power concerns), and clustered standard errors by listing. Below, we discuss results for the coefficient denoted by α_3, which indicated the difference in the effect of superhost status across levels of attributes.

Results from these models suggest that the effect of superhost status is less negative for listings from hosts with multiple listings ($\beta_{Superhost \times Multi}$ = .028, t(1,524,298) = 6.771, p < .001, 95% CI = [.020, .036]), and for more expensive listings ($\beta_{Superhost \times logPrice}$ = .007, t(1,523,862) = 2.306, p = .021, 95% CI = [0.001, .013]). This is perhaps surprising, as price may be expected to affect expectations in the same way as superhost status. However, it is likely that increases in prices reflect improvements in quality, and prices are significantly higher within-listing when listings are superhosts ($\beta_{Superhost}$ = 1.464, t(1,523,864) = 2.372, p = .018, 95% CI = [0.255, 2.674]).

We do not find differing effects of superhost status across physical attributes of listings. Specifically, there is no interaction of superhost status and the number of amenities listed (*M*_{Amenities} = 36.422, *SD* = 12.809; ($\beta_{Superhost \times Amenities}$ = –.0002, t(1,524,298) = 1.329, p = .184). There is also no interaction with the number of guests accommodated ($M_{\text{Accomodates}}$ = 4.521, *SD* = 2.899; ($\beta_{Superhost \times Accomodates}$ = .0003, t(1,524,298) = .389, p = .697). Airbnb also offers listings that are entire homes, as well as private rooms in hosts' homes, hotel rooms, and shared rooms. We did not find a difference in the effect of superhost status between entire homes and other listings ($\beta_{Superhost \times Entire}$ = .005, t(1,524,298) = .870, p = .384). This result complements our difference-in-differences with Vrbo, as all Vrbo listings are entire homes.

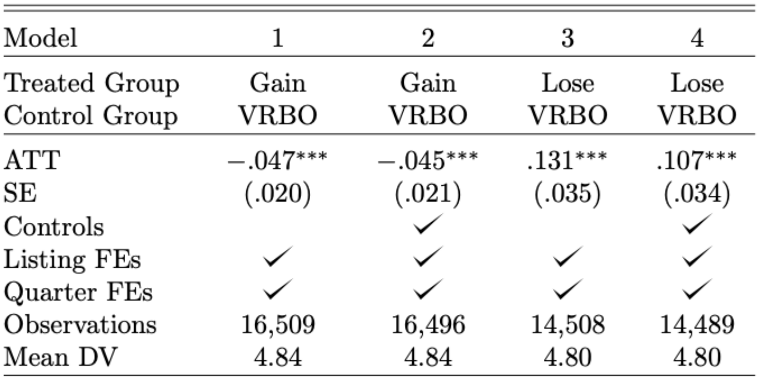
Finally, we test the robustness of the negative main effect of superhost status not across attributes, but across listings with different changes in status over time (i.e., gain only, lose only, both). Rather than presenting the interaction of superhost status with group, we replicated Model 3 from Table 5 among the three distinct subsets of listings with variation in status, and present those simple effects. These results suggest that superhost status is most negative for listings who only lose superhost status (Model 3 from Table 5; $\beta_{Superhost}$ = –.091, t(129,484) = –13.89, 95% CI = [–.104, –.078]) and both gain and lose status (Model 3 from Table 5; $\beta_{Superhost}$ = –.086, t(133,196) = –24.000, 95% CI = [–.093, –.079]), but less negative for those who only gain status (Model 3 from Table 5; $\beta_{Superhost}$ = –.047, t(167,115) = –11.729, 95% CI = [–.055, –.039]). This result is somewhat unexpected. Consumers only see if a listing is a superhost or not, and not a listing's former status. Therefore, it is not possible for consumers to pick up on these changes. Thus, these differences in estimates likely represent changes in listings over time, which we control for in the final identification by comparing ratings for listings to themselves on Vrbo.

WEB APPENDIX G: HETEROGENEITY OF AIRBNB-VRBO DIFFERENCE IN DIFFERENCES ANALYSIS

Because our Airbnb-Vrbo difference-in-differences strategy is the most causally defensible test of the effect of certification on ratings, we also performed a series of exploratory analyses on the heterogeneity of this effect. In contrast to Web Appendix F—which investigates heterogeneity through interactions of superhost status and property characteristics—the analyses herein consider subsets of the total data, as the estimation strategy we use (Callaway and Sant'Anna 2021) does not allow for interactions with covariates. Instead, we subsetted our data according to listing attributes one-at-a-time. For reference, we reproduce Table 6 from the main text, showing the main result of this analysis.

TABLE G1

RESULTS OF DIFFERENCE-IN-DIFFERENCES BETWEEN AIRBNB AND VRBO RATINGS FOR GAINING AND LOSING SUPERHOST STATUS



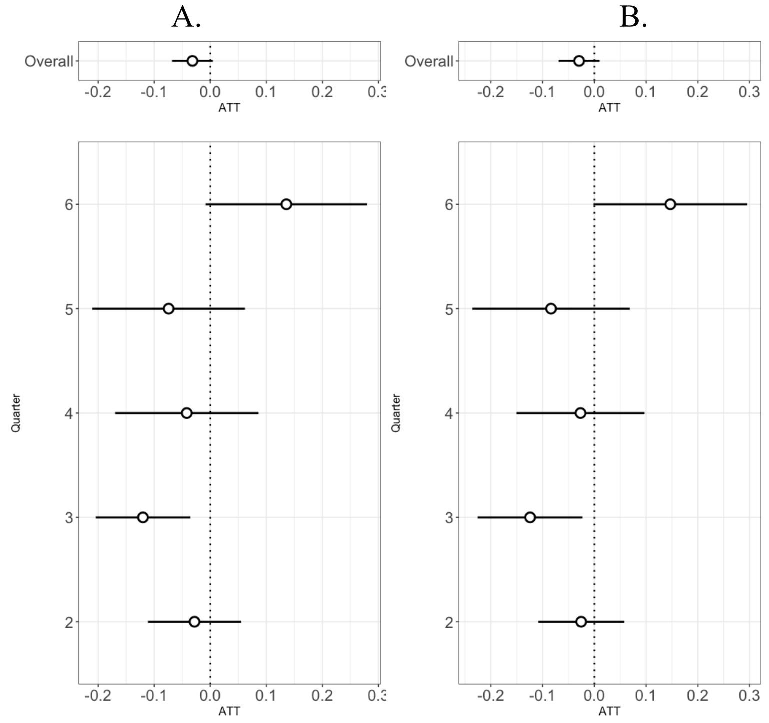
Model	1	2	3	4
Treated Group	Gain	Gain	Lose	Lose
Control Group	VRBO	VRBO	VRBO	VRBO
ATT	$-.047^{***}$	$-.045^{***}$	$.131^{***}$	$.107^{***}$
SE	(.020)	(.021)	(.035)	(.034)
Controls		✓		✓
Listing FEs	✓	✓	✓	✓
Quarter FEs	✓	✓	✓	✓
Observations	16,509	16,496	14,508	14,489
Mean DV	4.84	4.84	4.80	4.80

Disaggregated Across Quarters

First, we analyze the estimated average treatment on treated for superhost status between Airbnb and Vrbo at each time period individually, presenting these results in Figure G1 for listings who gain status (panel A without controls, panel B with controls) and Figure G2 for listings who gain status (panel A without controls, panel B with controls). Each point on each plot is the estimated difference-in-differences between Airbnb and Vrbo ratings among listings who change status in that quarter. Bars represent 95% confidence intervals.

FIGURE G1

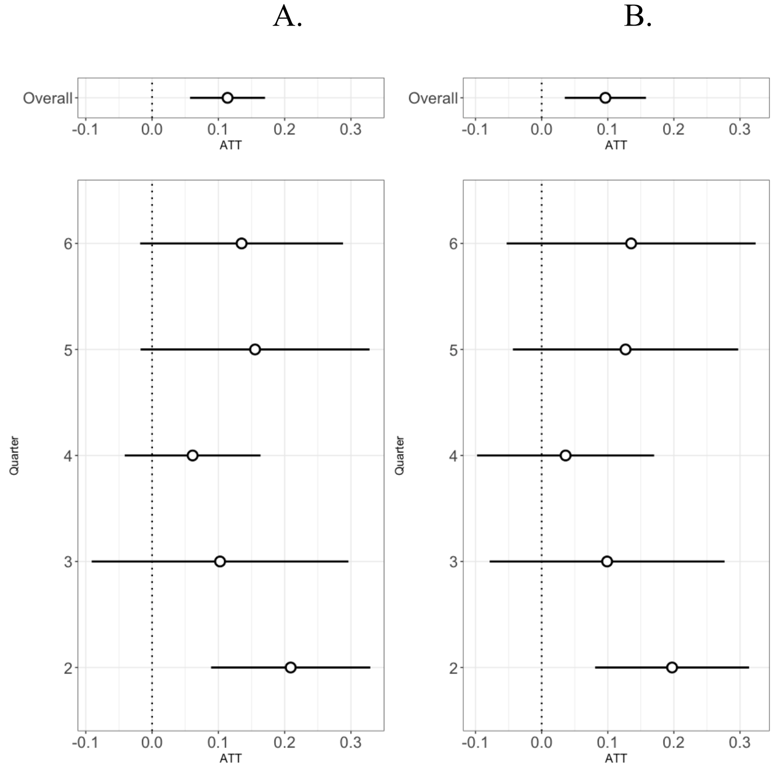
AVERAGE TREATMENT ON TREATED FOR AIRBNB LISTINGS WHO GAIN SUPERHOST STATUS DISAGGREGATED BY INDIVIDUAL TIME PERIODS



NOTE.–Panel A includes results from models without controls, Panel B includes results from models with controls.

FIGURE G2

AVERAGE TREATMENT ON TREATED FOR AIRBNB LISTINGS WHO LOSE SUPERHOST STATUS DISAGGREGATED BY INDIVIDUAL TIME PERIODS



NOTE.–Panel A includes results from models without controls, Panel B includes results from models with controls.

Among listings who gain status, four of five quarters demonstrate a negative effect. Surprisingly, the most recent quarter shows the opposite—a significant positive effect, such that ratings were significantly higher on Airbnb (compared to Vrbo) after gaining Airbnb superhost status. We do not have a theoretical explanation of this, but caution interpretation as this result relies on just 2,068 ratings between the two platforms—1,725 from Airbnb and 343 from Vrbo. Among listings who lose status, all five quarters demonstrate a positive effect of consistent size.

Analysis of Subsets of Listings

As with the within-listing Airbnb analysis, we next investigated heterogeneity between different kinds of listings. First, we estimated the ATT among listings who only changed status once—following from our estimation of regression to the mean in Web Appendix E. Next, we estimated the ATT among listings from hosts with multiple listings, and hosts with only a single listing. Then, we subset the data by a series of median splits on attributes of the listings. This includes average price, the amount by which price is increased when the listing is a superhost,[4] the number of guests accommodated, and the number of amenities listed. We estimate each difference once without controls, and once with controls for price, hosts' number of listings (where applicable), and number of ratings.

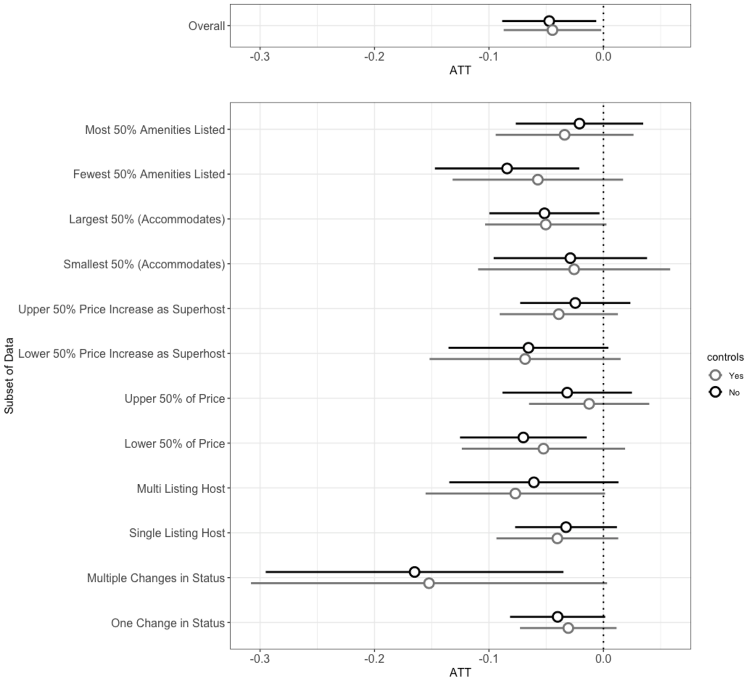
We present results for those who gain superhost status in Figure G3, where each point represents the estimated ATT, and bars represent the 95% confidence intervals. This finds consistent effects among each subpopulation, but with a larger negative effect for listings who change status multiple times, although these estimates only include observations before and after their first change.

We present results for those who lose superhost status in Figure G4, where each point represents the estimated ATT, and bars represent the 95% confidence intervals. This finds consistent effects among each subset, but with a less positive effect for listings whose hosts have multiple listings.

[4] Note: We median-split this separately for listings who gain and lose status, due to asymmetries between groups.

FIGURE G3

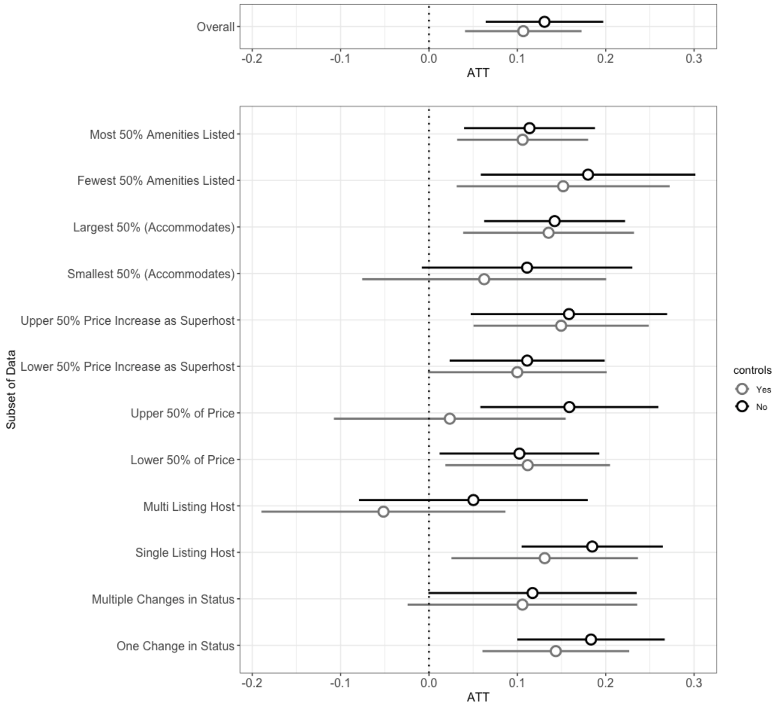
HETEROGENEITY OF AVERAGE TREATMENT ON TREATED FOR AIRBNB LISTINGS WHO GAIN SUPERHOST STATUS, COMPARED TO VRBO



NOTE.—This figure presents the average treatment effect on treated (ATT), first for the entire population of listings who gain status ("Overall"), and then for each subpopulation according to property characteristics. Points represent the ATT estimate, while extending lines represent 95% confidence intervals. Black points and lines represent estimates from models without controls, while grey points and lines represent estimates from models controlling for price, hosts' number of listings (where applicable), and number of ratings.

FIGURE G4

HETEROGENEITY OF AVERAGE TREATMENT ON TREATED FOR AIRBNB LISTINGS WHO LOSE SUPERHOST STATUS, COMPARED TO VRBO



NOTE.—This figure presents the average treatment effect on treated (ATT), first for the entire population of listings who lose status (“Overall”), and then for each subpopulation according to property characteristics. Points represent the ATT estimate, while extending lines represent 95% confidence intervals. Black points and lines represent estimates from models without controls, while grey points and lines represent estimates from models controlling for price, hosts’ number of listings (where applicable), and number of ratings.

WEB APPENDIX H: STUDY 3 ROBUSTNESS

Results Not Controlling for Order

The average quality rating is significantly different from the scale midpoint of 2.5 (*M* = 2.83; *t*(1,987) = 9.061; *p* < .001), with participants thinking the non-superhost (higher-rated) listing is of higher quality. Consistent results were observed for the choice dependent measure. The average participant was more likely to select the non-superhost (higher-rated) listing (*M* = .19; *t*(1,987) = 8.324; *p* < .001). In total, 54.93% of participants chose to stay with the non-superhost-tagged listing, compared to 36.12% for the superhost, and 8.95% indicating no preference.

Results Within Individual Cities

Los Angeles, California

For the first dependent measure, we found that quality perceptions differed across the order in which the superhost was presented (A or B; M_A = 2.82, M_B = 2.97; *t*(495) = 1.136; *p* = .257). Therefore, we do not collapse across this factor, although results are the same if we do. Controlling for the order of the superhost listing in the table, the average quality rating is significantly different from the scale midpoint of 2.5 (*M* = 2.89; *t*(495) = 5.864; *p* < .001), with participants thinking the non-superhost (higher-rated) listing is of higher quality. Consistent results were observed for the choice dependent measure. Controlling for the order the superhost appeared, the average participant was more likely to select the non-superhost (higher-rated) listing (*M* = .22; *t*(495) = 5.332; *p* < .001). In total, 56.34% of participants chose to stay with the non-superhost-tagged listing, compared to 34.21% for the superhost, and 9.46% indicating no preference.

San Francisco, California

For the first dependent measure, we found that quality perceptions differed across the order in which the superhost was presented (A or B; M_A = 3.06, M_B = 3.02; t(495) = –.272; p = .786). Therefore, we do not collapse across this factor, although results are the same if we do. Controlling for the order of the superhost listing in the table, the average quality rating is significantly different from the scale midpoint of 2.5 (M = 3.04; t(495) = 8.026; p < .001), with participants thinking the non-superhost (higher-rated) listing is of higher quality. Consistent results were observed for the choice dependent measure. Controlling for the order the superhost appeared, the average participant was more likely to select the non-superhost (higher-rated) listing (M = .31; t(495) = 7.761; p < .001). In total, 61.37% of participants chose to stay with the non-superhost-tagged listing, compared to 29.98% for the superhost, and 8.65% indicating no preference.

Niagara Falls, New York

For the first dependent measure, we found that quality perceptions differed across the order in which the superhost was presented (A or B; M_A = 2.43, M_B = 2.55; t(495) = .974; p = .331). Therefore, we do not collapse across this factor, although results are the same if we do. Controlling for the order of the superhost listing in the table, the average quality rating is not significantly different from the scale midpoint of 2.5 (M = 2.49; t(495) = –.149; p = .881). Consistent results were observed for the choice dependent measure. Controlling for the order the superhost appeared, the average participant was not more likely to select the non-superhost (higher-rated) listing (M = –.02; t(495) = –.468; p = .64). In total, 44.47% of participants chose to stay with the non-superhost-tagged listing, compared to 46.48% for the superhost, and 9.05% indicating no preference.

Moab, Utah

For the first dependent measure, we found that quality perceptions differed across the order in which the superhost was presented (A or B; $M_A = 2.76$, $M_B = 3.04$; $t(495) = 2.196$; $p = .029$). Therefore, we do not collapse across this factor, although results are the same if we do. Controlling for the order of the superhost listing in the table, the average quality rating is significantly different from the scale midpoint of 2.5 ($M = 2.9$; $t(495) = 6.272$; $p < .001$), with participants thinking the non-superhost (higher-rated) listing is of higher quality. Consistent results were observed for the choice dependent measure. Controlling for the order the superhost appeared, the average participant was more likely to select the non-superhost (higher-rated) listing ($M = .24$; $t(495) = 5.705$; $p < .001$). In total, 57.55% of participants chose to stay with the non-superhost-tagged listing, compared to 33.8% for the superhost, and 8.65% indicating no preference.