More Stars or More Reviews? Differential Effects of Reputation on Trust in the Sharing Economy

Will Qiu
Stanford University
Stanford, CA, USA
willqiu@stanford.edu

Palo Parigi
Airbnb Research, Airbnb
San Francisco, CA, USA
Stanford University
Stanford, CA, USA
pparigi@stanford.edu

Bruno Abrahao
NYU Shanghai
Shanghai, China
Stanford University
Stanford, CA, USA
abrahao@cs.stanford.edu

ABSTRACT

The large majority of reputation systems use features such as star ratings and reviews to give users a reputation in online peer-to-peer markets. Both have been shown to be effective for signaling trustworthiness. However, the exact extent to which these features can change perceptions of users' trustworthiness remains an open question. Using data from an online experiment conducted on Airbnb users, we investigate which of the two types of reputation information –average star rating or the number of reviews –is more important for signaling a user's trustworthiness. We find that the relative effectiveness of ratings and reviews differ depending on whether reputation has a strong or a weak differentiation power. Our findings show that reputation effects are contingent on and susceptible to the context created by the alternative choices presented to users, highlighting how reputation information is displayed can drastically alter their efficacy for engendering trust.

ACM Classification Keywords

H.5.m Information Interfaces and Presentation (e.g., HCI): Miscellaneous

Author Keywords

Reputation and Rating Systems; Trust; Sharing Economy; Airbnb

INTRODUCTION

The use of sharing economy services has exploded in recent years. Sharing platforms such as Airbnb and Uber predicate their business models upon connecting individuals interested in the exchange of a good or service. While the popularity of these services has increased dramatically, trust nevertheless remains a salient issue on these platforms [24]. Asking users to place trust in a total stranger can be difficult because of the associated risks involved. Scholars have found that users tend to have varying propensities to do so depending on both the type of platform they are using and their social background. [11, 25]

By far the most ubiquitously adopted mechanism by sharing platforms to increase trust among their users during exchange is the online reputation system [26]. A decision which influenced primarily by traditional peer-to-peer (p2p) markets such as eBay [21]. In most cases, sharing platforms have implemented the conventional 5-star rating system. This system is also usually accompanied by a user review mechanism. The star rating is the simple average rating of the entire history a user's activity on the platform and reviews provide additional unstructured information regarding each exchange. Recent work has shown that this system can be quite effective for signalling users' trustworthiness by reducing social bias [1]. What remains to be explored, however, is the individual effects of star ratings and reviews for influencing these perceptions. Indeed, the question remains as to whether these two separate dimensions of reputation have similar effects and if such effects are constant under various conditions.

The goal of this article is to understand the relative efficacy of star ratings and reviews for signaling trustworthiness by using an online experiment designed specifically to measure trust between sharing economy users. To this end, we designed and conducted a large-scale online experiment using participants recruited from the popular hospitality platform Airbnb. The experiment takes the form of an "investment" game, whereby subjects are shown a fixed number of profiles of other Airbnb members and made credit allocations between them. Each of these profiles contains information about a given individual, including their average start rating and review counts on the platform. We measure how subjects' perceptions of the trustworthiness of a given profile is influenced by their exposure to various reputation information, by randomly assigning them into different reputation conditions where we varied the two main features of the standard reputation system described above –star ratings and review counts.

Our findings show that the relative effects of stars versus reviews is contingent on the underlying context in which this information is presented to subjects. Specifically, in contexts where reputation serves as a strong enough signal for differentiating between choices, having many reviews (versus only having a few) is much more valuable for signaling trustworthiness than having a five star rating (versus having only four).

CHI 2018, April 21–26, 2018, Montréal, QC, Canada.

ACM ISBN 978-1-4503-5620-6/18/04 ...$15.00
https://doi.org/10.1145/3173574.3173727

However, under conditions where the differentiating power of reputation is low, the relative effects of stars and review counts are about the same. This finding highlights that reputation signals are highly malleable to contextual factors that affect their effectiveness for helping users differentiate quality. Our findings here thus have practical design implications for sharing economy platforms that rely on star ratings and reviews as the main reputation mechanisms for facilitating exchange between their users.

BACKGROUND AND RELATED WORK

Reputation systems in online markets

An extensive literature has argued that online reputation systems are highly conducive for engendering trust between participants in online markets [22, 27]. The basic theoretical claims are straightforward. Reputation systems allow users to share their past exchange experiences publicly by rating them. Others then use this information to form expectations about how trustworthy a particular user would likely be. On the other end, reputation also acts as an incentive for users to be trustworthy within the marketplace in order to ensure future exchange opportunities [16]. Online reputation thus proves to be useful for sharing economy platforms specifically because knowing how others have rated a person's past interactions on the platform provides users information about his or her future propensities to act in a trustworthy manner. This means that users with higher reputations will generally be viewed as more trustworthy by others than their lower reputation peers.

Empirical studies of reputation systems in peer-to-peer markets both within the lab [6] as well as in actual online markets [23, 20, 18, 10, 22] have broadly supported the theoretical claims about their positive effects for facilitating exchange. In general, these studies have found a positive association between reputation and trust –usually indicated by the fact that high reputation sellers are able to sell their products at a premium. Although these initial studies are informative, they are also somewhat ambiguous and inconsistent with respect to quantifying the exact effects and the relative efficacy of these mechanisms for engendering trust [7]. Furthermore, they do not consider in detail how different features of a users' online reputation –for example, a user's star rating and reviews –can have different effects on perceptions of trustworthiness. Assuming that disparate components of a reputation system tend to measure different dimensions of reputation, then understanding how they affect perceptions of trustworthiness differently would provide significant value to platform designers.

Reputation system design

Some recent work that have tried to pinpoint the exact effects of various reputation systems on trust in online markets have stressed the idea that their observed effects could be subject to change depending on how a particular system is designed. This group of studies has suggested that factors such as (1) the underlying metric by which reputation information is displayed [8], (2) the availability and the rate of update of reputation information [9], and (3) the strategic interaction between individuals in the production of reputation [5] all play a role in the underlying dynamics and efficacy of reputation signals.

Reputation system design in HCI

While the aforementioned work have highlighted the importance of design for affecting the effectiveness of reputation signals, their claims have not been tested using experiments that allow for causal inference. Furthermore, these studies have examined reputation system design from an economic perspective and not from the perspective of user experience, relying largely on formal models. Indeed, some preliminary work has shown that basic front-end features displayed on a user profile can have significant consequences for trust-based judgments that users make on sharing platforms. Researchers have found, for example, that information such as users' number of friends and profile views [17], self-disclosures and descriptions [19], and personal photos [12] all play a significant role in signalling their trustworthiness to others.

Thus, understanding how reviews and ratings influence user behavior as front-end features is instrumental for designing a more effective reputation system. Related to this, Ert *et al.* [12] uncovered an important auxiliary finding that serves as the basis for our current study. Specifically, in their study of host photos they found that when subjects were asked to rate their choice preferences among a group of profiles that had a more compressed distribution of ratings (specifically, between 4.5 stars and 5 stars), rating effects became nonexistent compared to the situation in which they are presented with profiles whose ratings occupied a wider distribution (between 3.5 stars and 5 stars). The findings further suggests the possibility that users are susceptible to reference dependence when making decisions –a principle popularized by behavioral economics [15]. Fields like marketing have long used this idea to study various phenomena such as reference pricing [3].

Applied to reputation systems, this general theoretical framework would argue that users do not evaluate reputation signals independently, but rather form subjective assessments of the underlying value of a given piece of reputation information with respect to some reference point. They evaluate a given profile and its associated reputation in contrast to the surrounding context of other concurrent available reputation information. These available alternatives form an underlying distribution of reputation levels. When such a distribution is "compressed", the differentiating power of an individual reputation is diminished if it is situated close to the reference, making it difficult for users to choose among all available alternatives. In this situation, profiles with reputation levels that prove to deviate significantly from the reference would then be evaluated much more strongly due to their large contrast relative to the baseline. If such processes are present, then they provide one of the first demonstrations of potential unintended consequences of "reputation inflation" that have been shown to exist in many online markets, whereby reputation signals would not induce user choice appropriately, if at all [28, 13, 14]. Our experiment here thus provides a more systematic attempt to uncover whether this hypothesized set of effects exist for ratings and reviews and whether their relative effects under these set of conditions are similar.

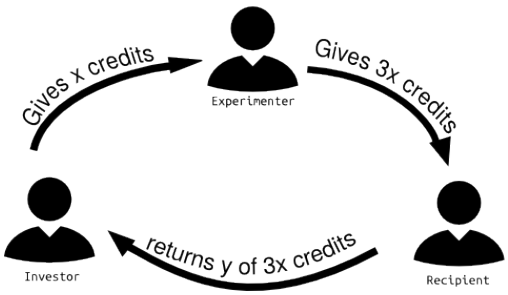

Figure 1. One iteration of the standard trust game. Investor gives x number of credits, the experimenter triples it before handing it to the recipient, and the recipient then returns y of the $3x$ amount he/she received.

EXPERIMENTAL DESIGN

Method

To investigate the questions posed above we conducted an online experiment in collaboration with Airbnb.The data used in this paper was designed by authors of earlier work that have also used it to investigate trust in the sharing economy [1]. An e-mail invitation to participate in an online study for a monetary reward was sent to 100,000 U.S. Airbnb users. Of these, 8,906 registered and 5,277 completed the experiment.[1]. Participants were directed to an experiment platform where they played a trust game after providing some basic demographic information indicating their age, gender, marital status, and home state (zip code).

The trust game

Participants were invited to play a variation of the widely studied trust game with five other Airbnb profiles. This game was originally developed by economists in an effort to measure trust in economic exchange within the context of a laboratory setting [2]. In the original version of the game, participants are anonymously paired together and asked to make a series of exchanges using a set of credits. Each individual is assigned to play either the role of an "investor" or a "recipient". A single round of the game consists of an investment phase and a return phase. During the investment phase, the investor first decides how much of his/her allotted credits to give to the recipient. The amount that he/she sends over is then usually tripled by the experimenter before it is sent to the recipient. In the return round, the recipient decides how much of the credits he/she received (now 3x the original amount) back to the investor. Performance in the game is measured by the amount of credits that a participant has accumulated at the end of the game. Participants are usually paid a monetary reward commensurate with their final credit count to act as an incentive for them to make the best "investment decisions". Figure 1 depicts one round of play in the trust game.

The validity of the trust game for measuring trust is predicated upon the fact that under standard economic theory, rational investors would never invest any credits in the recipient. Since investors and recipients do not know each other ahead of time, investors should not have any expectation that *any* credits they give away will be returned. Instead, investors would expect that rationally acting recipients would choose to keep all of their credits. As such, the predicted Nash equilibrium for this game would be that the investor always sends 0 credits, and the recipient always returns 0 credits [2]. Any empirical deviation from this prediction –i.e., if the investor gives a credit amount greater than 0 to the recipient –can then be considered as a measure of the amount of "trust" that he or she has in the recipient, which can also be taken as a measure of how trustworthy the investor perceives the recipient to be.[2].

For our current study, we made a number of modifications to the standard trust game described above. This was done to reflect as closely as possible the nature of trust-based decision-making on Airbnb –namely, the initial selection of exchange partners on the platform. First, unlike the traditional trust game –which usually consists of multiple rounds –participants only played a single round of investment. Second, in our experiment, every participant was assigned the role of the investor, but were told that some were chosen to be recipients. When the game begins, they are paired with a selection of five profiles presented to them as other randomly chosen Airbnb users that were really randomly generated profiles. They then invested x amount of the 100 total credits given to them at the beginning (see Figure 2).

Every recipient profile contained a combination of demographic as well as reputation attributes, which were randomly generated according to a set of prescribed rules. This information was made to reflect the type of guest and host information that Airbnb users would typically be privy to when searching for listings on the platform. As incentive for participants to make the *best* choices as possible (investments that they believe will net them the largest return at the end of the game), we tied their performance in the game to the potential for winning a monetary reward. Specifically, the top 100 participants with the largest amount of credits accumulated won a $100 dollar gift card. [3]

Experimental conditions

While the overall experiment itself employs a mixed design, our analysis of the reputation conditions here is strictly cross-subject. Below we briefly describe both the demographic as well as the reputation conditions in the full experiment in order to facilitate a better understanding of our analysis. We briefly provide some rationale of why such an experimental design is appropriate for our analysis below, a more detailed explanation and justification of the entire experiment can be found in [1].

[1]The rest of the registered users did not complete all phases required for the present study

[2]Trust here implies that the investor has an expectation that some amount of the tripled credits will be returned back to him/her.

[3]Since participants are playing against artificial profiles, the returns that they received for a given "investment portfolio" had to be simulated. We derived the probability of winning with a given portfolio using an empirical distribution of recipient credit returns obtained from the results of a previous experiment we conducted where participants played as both investors and recipients.

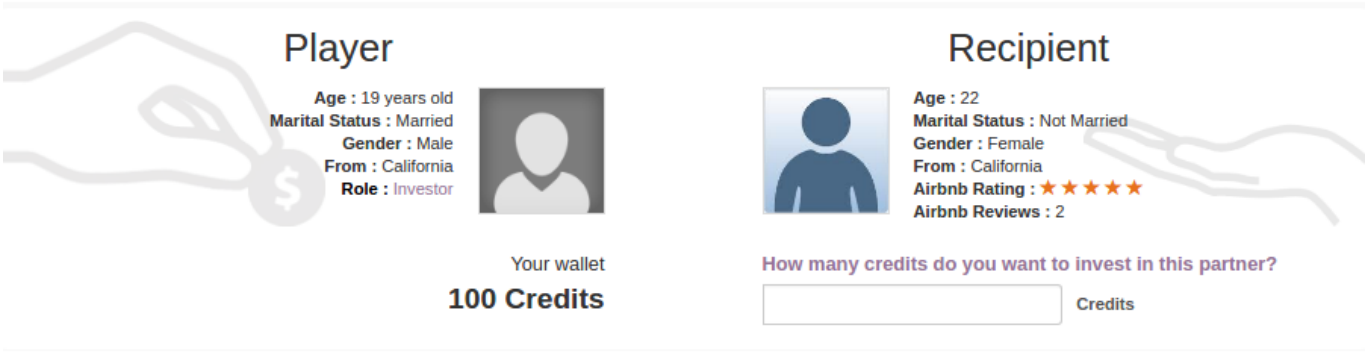

Figure 2. Each subject sees five of these recipient profiles during the game, and must decide how much of their credits to invest in each. Each profile displays the recipient's social characteristics as well as reputation level on Airbnb. This particular profile has a social distance of 2 from the player. They differ in their gender and marital status.

Socio-demographic conditions

For generating the socio-demographic characteristics of the profiles, we applied Blau's [4] concept of the social space to create profiles according to a uni-dimensional measure of social distance d to participants. Here, d ranges from 0 (complete similarity between a participant and a profile) to 4 (complete social dissimilarity). Every participant sees on a single page in random order one profile at each level of social distance of $d = \{0, 1, 2\}$. They saw two profiles with the social distance of $d = 4$. As it will be made clear, this paper focuses on the difference between these two most socially distant profiles. [4]

Reputation conditions

To motivate the main part of our analysis here, profiles were also designed to contain two pieces of reputation information: an average star rating and a count of the number of reviews they have on Airbnb. These two dimensions of reputation were manipulated and various combinations of their respective levels were made in order to form specific cross-subject conditions which participants were randomly assigned to.

Creating reputational "worlds"

At the broadest level, each set of profiles that a participant encounters occupy one of two possible reputational contexts –which we gave the label of "worlds". In both, we set the first four profiles at social distances $d = \{0, 1, 2, 4\}$ to share the same star rating and review count level, followed by selecting one of the two reputation features –either star rating or review count level –of the second distance $d = 4$ profile to be either higher or lower than that of the first four.[5] Table 1 displays this general setup. For simplicity, we refer to the distance $d = 4$ profile with the same reputation as the other three profiles as *p4* and the other distance $d = 4$ profile with a reputation different from the rest as *p5*). We label the other profiles by their assigned social distance d to the participant (*p0, p1, p2*)

[4]We did not present participants with a profile at a social distance of $d = 3$. The reason is mostly practical, namely to to minimize attrition from participant fatigue while still being able to measure the effects of social distance.

[5]The reputation feature that was not selected is fixed to the same level as the other four profiles.

	Social Distance	Reputation Level
Profile 0 (*p0*)	0	same
Profile 1 (*p1*)	1	same
Profile 2 (*p2*)	2	same
Profile 4 (*p4*)	4	same
Profile 5 (*p5*)	4	different

Table 1. General setup of social and reputation conditions of recipient profiles across all conditions. The social distances are set up are uniform across subjects. But reputation condition differ across subjects with respect to the direction and size of the difference between profile 5 (loner profile) and the other 4 profiles.

The two worlds we created differ primarily with respect to the underlying reputational *differentiation* that profiles signal to participants. In world one, *p5* will have an overall lower reputation level than the other four profiles. For example, if the other four profiles have 5 stars, then *p5* will have a star rating of either 4 or 0. Inversely, in world two, *p5* will have an overall reputation that is higher than the other four profiles (along either the dimension of stars or review counts). World one thus measures the effect of social distance on trust without reputation counterbalancing the general tendency toward placing trust on strangers that look familiar, i.e., located at closer distances. On the other hand, world two introduces reputation as a potential counter force toward familiarity. Conceptually, world one creates an environment where reputation has a weak differentiating power, because the *highest* reputation is shared by most profiles (from *p0* to *p4*) and thus it is difficult to distinguish among alternatives using reputation information. In contrast, world two creates a context where reputation has a *strong* differentiating power, because profile *p5* has the best reputation, and thus should present to participants as the clear best choice.

Within-world reputation conditions

Within each world we also constructed a number of cross-subject reputation conditions that differ primarily with respect to the substantive reputation levels that the profiles have. To simplify the number of possible conditions we had to generate

Condition	*p4* stars	*p5* stars	*p4* reviews	*p5* reviews	World	n
1A	5	4	High	High	One	669
1B	5	4	Low	Low	One	423
1C	5	5	High	Low	One	666
1D	4	4	High	Low	One	450
2A	4	5	High	High	Two	612
2B	4	5	Low	Low	Two	433
2C	5	5	Low	High	Two	694
2D	4	4	Low	High	Two	433

Table 2. Counts of subjects across eight unique reputation conditions constructed across subjects. "World" refers to one of two possible contexts where profile 5 (*p5*) will have a different reputation level as the rest of the profiles. For review count levels, Low = 1-3 reviews, High = 10-100 reviews. We did not include here other conditions where profiles had 0 stars and 0 reviews since these conditions were not part of our final analysis.

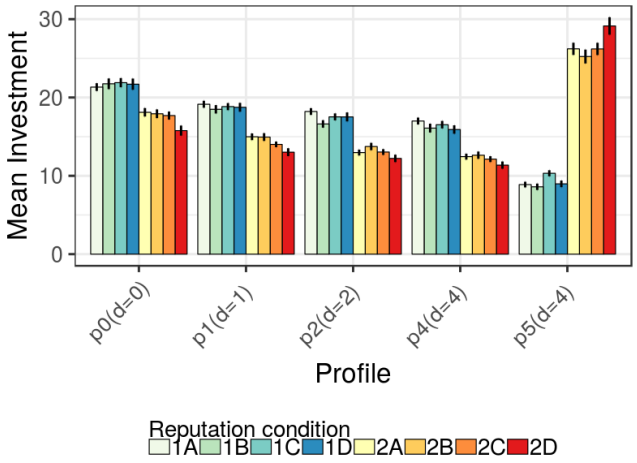

Figure 3. Barplot of mean investment of all five profiles (*p0 - p5*) by condition. *d* refers to social distance of a profile to subjects. Black lines indicate standard error of the mean

to achieve a viable sample size for each, we limited the possible star ratings to either 0 stars, 4 stars, or 5 stars. Likewise, we discretized review counts into three different levels: None (0 reviews), Low (1-3 reviews), and High (11 - 50 reviews). Both star ratings and review count levels were allowed to independently vary with the exception of the corner case of 0 stars and 0 reviews, which could only logically appear in tandem.[6]

Table 2 lists the different cross-subject reputation conditions that were constructed by the *p4* and *p5* star and review levels and world. The last column lists the sample sizes that were obtained. These conditions are symmetric, meaning that each individually lettered conditions within a given world has a mirrored counterpart in the other world whereby reputation levels of *p4* and *p5* are swapped. We list only the reputation levels of *p4* and *p5* here since all other profiles (*p0*, *p1*, and *p2*) share the same reputation as *p4*. Note that in each condition only one dimension of reputation (either star ratings or review count but not both) is varied. This ensures that the differences in credit investments made by participants between two given profiles can be properly identified. Furthermore, we did not include conditions with the corner case of 0 stars and 0 reviews in our analysis, since separate effects of stars and reviews are not identifiable in those cases.

ANALYSIS

Our analysis proceeds in two parts. First, we estimate the simple average effects of each reputation feature on trustworthiness found in our experiment. Specifically, how does the perception of a profile's trustworthiness change when it goes from having an average rating of 4 stars to 5 stars and from an indicated review count of low (1-3) to a high (11-50)? Following this, we investigate how the context changes these observed effects. Specifically, how does the distribution of reputation information among profiles change how users interpret these signals? Do subjects evaluate the value of stars and reviews equally under contexts of low reputation differentiation (world one) versus contexts of high reputation differentiation (world two)?

Figure 3 plots the mean investments for each condition within each world separated by profile, and 4 plots the distribution of investments. We can see that when reputation is not a factor counterbalancing a general tendency toward trusting what seems familiar (world one), the most similar profile to the participant ($d = 0$) receives the greatest amount of investment and the most dissimilar profile ($d = 4$) receives the least. This holds true for all the conditions within world one. The situation is reversed in world two, when reputation counterbalanced the tendency toward familiarity.

In the current analysis we restrict our investigation to *p5* and *p4* investments only. A comprehensive analysis of these other profiles is presented in [1]. Here we are considering the two profiles that have the highest implications for HCI, which was beyond the scope of [1]. Restricting our analysis to these two profiles also allows us to remove social distance as a confounding factor. Since *p4* and *p5* share the exact same social distance to the participant, absent of reputation differences, each profile would be expected to receive similar credit amounts from participants. Any observed differences in the investments received between the two profiles would thus reasonably stem from factors that are not related to social distance.

What is the effect of star ratings and review counts on perceptions of trustworthiness?

We begin our analysis first by finding the simple average effect of star ratings and review counts on subjects' investment in profiles *p4* and *p5*. To do this, we fit a mixed multivariate regression model using profile investment amount as our dependent variable, and star rating and review counts as independent variables. Specifically, we fit the following model:

[6]This was needed to ensure a level of realism during the experiment since it makes little logical sense for a profile to have 0 stars but not have 0 reviews.

Table 3. Multi-level Model Estimates of Star Ratings and Review Counts Components on Investment

Covariate	Dependent variable:				
	Investment				
	(1) (Intercept Only)	(2) (Star Only)	(3) (Review Only)	(4) (No Interaction)	(5) (Full Model)
Intercept	16.09***(0.1609)	13.16***(0.237)	12.91***(0.235)	9.534***(0.292)	7.08***(0.3718)
Star = 5		5.285*** (0.3187)		5.69***(0.31)	4.497***(0.452)
Review = H			5.799***(0.317)	6.21***(0.311)	5.16***(0.45)
Star = 5 $\times$ Review = H					1.487(0.60)
World = two					6.53*** (0.303)
Variance Components					
Subject	0.0	7.56×10^{-12}	3.899×10^{-10}	14.06	1.58
Residual	226.7	219.8	218.4	197.34	198.94
Observations	8760	8760	8760	8760	8760
AIC	72378.0	72109.4	72052.2	71749.4	71309.8
BIC	72399.3	72137.7	72080.5	71784.8	71359.3
Log Likelihood	-36050.7	-40112.3	-36022.1	-35869.7	-35647.9

Note: Observations are from participant investments in p4 p5 in conditions listed in Table 2.

**p<0.05; ***p<0.01

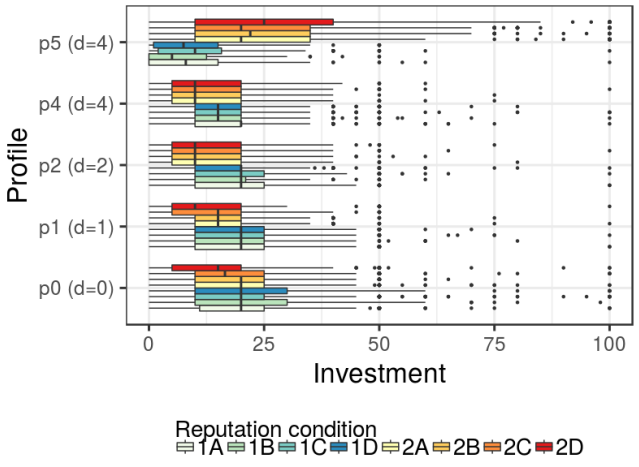

Figure 4. Distribution of investment in profiles (*p0 - p5*) by condition. *d* refers to social distance of a profile to subjects.

$$\hat{Y_{ij}} = \hat{\mu} + \hat{\alpha_j} + \hat{\beta_1} s_i + \hat{\beta_2} r_i + \hat{\beta_3}(s_i \times r_i) + \hat{\beta_4} w_i + e_{ij}$$

Where $\hat{Y_{ij}}$ is the predicted investment amount for profile i by subject j. $\hat{\mu}$ is the global intercept at star rating of 4 and Low review count in world one. $\hat{\beta_1}$ is the profile level estimate of having 5 stars (s), $\hat{\beta_2}$ is the profile level estimate of having High review counts (r) and $\hat{\beta_3}$ is their estimated interaction effect. $\hat{\beta_4}$ is the estimate of a profile i being placed in world two. e_{ij} is the random error of profile i per subject j. Star rating (s) is a factor variable with two levels (4 stars or 5 stars) and review condition (r) also has two levels (Low or High). Because there were multiple measurements of investments per participant (each participant made an investment in *p4* and in *p5*), the measured investments are correlated. To account for this we nested profile investments within subjects, by fitting simple random subject-level intercept α_j. The model also does not include an explicit term for social distance because our analysis is confined to observed investments between *p4* and *p5* only, who share the exact same distance to the subject.

Table 3 shows estimated fixed effect coefficients from our data. We estimated five different models: (1) an intercept only model, (2) star rating only model, (3) review count only model, (4) additive model of star and review, and lastly a full model with (5) both additive as well as interaction terms. The table also presents the variances of our subject-level intercept as well as the residual variances. In the intercept only model, the intercept is simply the global mean of the entire estimated

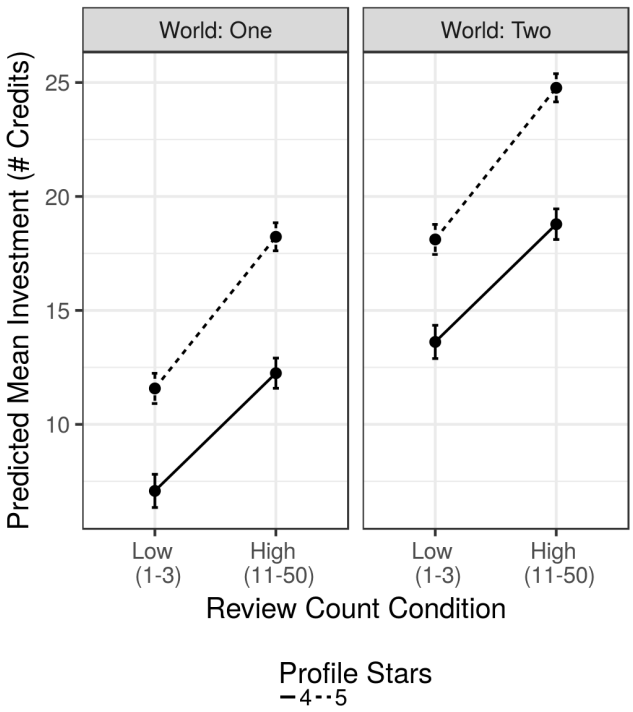

Figure 5. Predicted Investment from model estimates at each level of Stars and Review Counts, bars indicate 95% confidence intervals

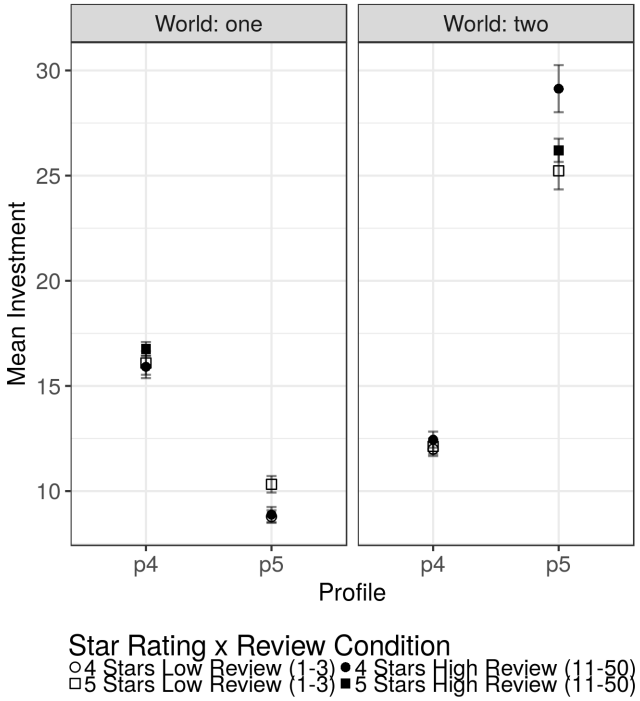

Figure 6. Mean investment in *p4* and *p5* by world, and star rating x review condition. Bars indicate between-subject standard errors

sample. In the next four models the intercept refers to the mean predicted investment at the reference level of each of the terms included. The estimated coefficients for each term can be interpreted as the average change in the amount of investment in a given profile as a profile reputation dimension changes from one the baseline level to the new level, net of any changes on the other reputation dimension.

We can see from the results that there is clearly a significant increase in average investment received when a profile goes from having 4 stars to 5 stars (~ 4.5 more credits or ~ 0.33 standard units) as well as going from low review to high review (~ 5.16 more credits or ~ 0.38 standard units). Their interaction (β_3) is not statistically significant. A linear hypothesis test of equality shows however that these two coefficients are not significantly different from each other ($\chi^2 = 2.44$, $p = 0.11$). These estimates are stable even when we take into account world differences. In other words, for a profile, the effect of going from having 4 stars to 5 stars on the amount of credits is equivalent to the effect of going from having only 1-3 reviews to having at least 11 reviews on average. Figure 5 plots the predicted average investment amount at different levels of star rating and review counts.

Rating and review effects under different reputational contexts

The previous analysis found a clear significant effect of having more stars and more reviews on how many credits a profile received from subjects. The model we estimated looked at the *average* effect of these two features on subjects' invested amounts and found that they are largely equivalent. However, recall that the experiment also placed subjects into distinctive reputational worlds, whereby the underlying distributions of reputation among the profiles that they encounter are systematically varied to have different baseline reputations. Specifically, in world one, *p5* will always have an overall reputation that is lower than the other four profiles (which share the same exact baseline reputation), whereas in world two *p5* will always have an overall reputation that is higher than the rest.

Conceptually, the two world can be characterized by their differences in the level of differentiation of reputations that exist between profiles. Reputation signals are weak in world one because a majority of the profiles (from *p0* to *p4*) share the same highest possible reputation among each other. As such, participants do not have a clear best option when making investments among the first four profiles based on reputation alone, leading them to defer to social distance as a heuristic. Indeed, we can see from the preliminary average investments in Figure 3 that across all world one conditions profile investments decrease consistently with social distance, with *p5* receive the least amount due to being both the furthest away from the subject *and* having the lowest reputation. On the other hand, we see in all world two conditions that *p5* receives a drastically higher amount of investment than the other four profiles. In these conditions, participants can easily identify the most ideal profile –*p5* –since it has the highest reputation.

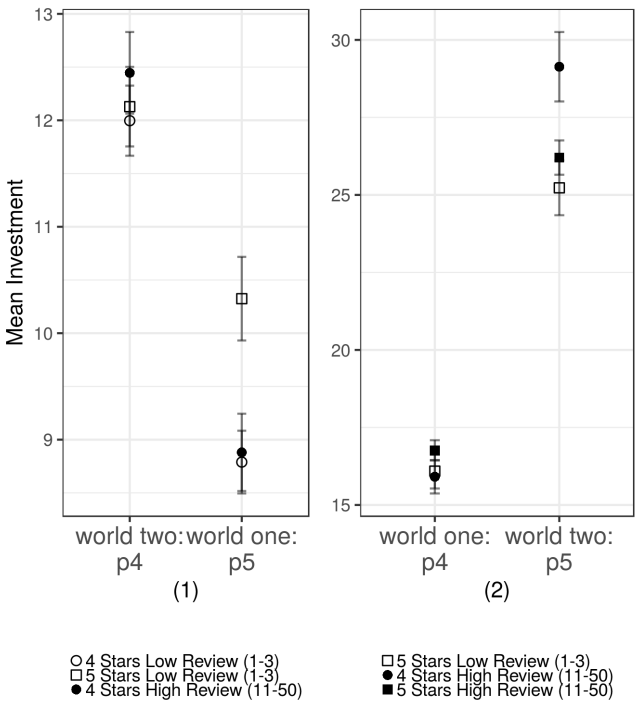

Figure 7. Mean investment in *p4* and *p5* by world, and star rating x review condition. Left panel compares means of *p4* in world two with *p5* in world one. Right panel compares *p4* in world one and *p5* in world two. Bars indicate between-subject standard errors. Difference between investment in each panel are the cross-world premiums on for each reputation category.

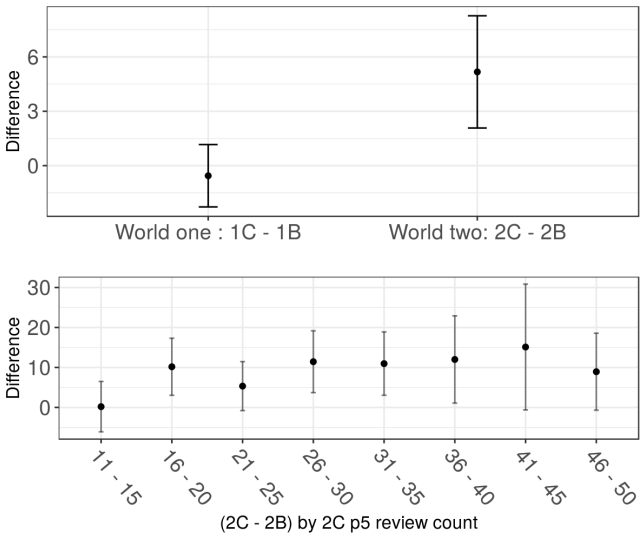

Figure 8. Plot of difference of mean differences between *p5* and *p4* investment in conditions 1C vs 1B and 2C vs 2B. Top graph plots the mean difference across all review counts. Bottom plot the mean differences of 2C and 2B across 2C *p5* review counts. Error bars are 95% confidence intervals, which takes into account the covariance between *p5* and *p4* investments.

Examining in more detail the mean investments for *p4* and *p5* between worlds highlights this effect further. Figure 7 plots the mean investment by subjects in *p5* and *p4*. The left panel (1) compares investments in *p4* in world two with investments in*p5* in world one, while the right panel (2) compares investments in world one *p4* to investments in *p5* in world two. In world two we can observe a clear positive difference in credits received by *p5* with 4 stars and high review counts and investments received by *p5* with 5 stars and low review counts across subjects –this difference goes away in world one (see Figure 6). Both of these observations suggest that participants assess trustworthiness differently depending on the differentiating power of the reputation.

In order to determine exactly how the two worlds affect how subjects assess trustworthiness, we investigate whether having 4 stars but a high review count on a profile serves as a better predictor of a subject's investment than having 5 stars with a low review count. Under our mixed model, we should expect that the 1 star effect is largely equivalent to an increase of having 1-3 reviews to having 11 or more reviews on one's profile. Thus, if there are differences between these two reputational profiles between worlds, we would then expect world effects to be in play.

Figure 7 shows instead their effects on credit received from participants are not the same. Specifically, we find that in world two, a *p5* that had 4 stars but a high review count (11+) had more credits invested by subjects than *p5* that had 5 stars with a low review count. This shows that in a context where reputation has high differentiating power (i.e., world two), number of reviews drive trustworthiness more so than ratings.

How big is the review count premium over 1 star?

Due to the way we constructed the reputation conditions during the experiment, we can actually quantify the exact amount of the credit premium that reviews have over stars. To do this, we construct a new dependent measure using the raw investment counts of *p4* and *p5* for each subject. Specifically, we calculate the *difference* between a subject *i*'s investment in profile 5 (*p5*) and profile 4 (*p4*):

$$y_i^* = Invest_i^{p5} - Invest_i^{p4}$$

Since *p4* and *p5* are identical with respect to their social distance to the subject, absent of any reputation differences, the expected value of y_i^* would be 0. Table 4 lists all the relevant reputation conditions sorted by world and profile star $\times$ review counts. The last column of the table describes the exact type of effect measured when y_i^* is calculated for that condition. We can see that since in world two both *p5* reputation levels of 4H and 5L conditions are compared with *p4* reputation of 4L (conditions 2B and 2C), we can directly compare cross-condition differences of the within-subject profile investment differences. For example, the investment difference between *p5* and *p4* in 2B measures the change in 1 star difference (5L - 4L), and in 2C this difference measures the change in review levels (4H - 4L), subtracting one from the other (2C - 2B) gives us the average difference between the 1 star and review level

Condition	World	Profile 5 (*p5*) stars × review count	Profile 4 (*p4*) star × review count	Measured effect of invest^{p5} − invest^{p4}
1A	one	4 Stars High Review (4H)	5 Stars High Review (5H)	1 star change
1B	one	4 Stars Low Review (4L)	5 Stars Low Review (5L)	1 star change
1C	one	4 Stars Low Review (4L)	4 Stars High Review (4H)	review level change
1D	one	5 Stars Low Review (5L)	5 Stars High Review (5H)	review level change
2A	two	5 Stars High Review (5H)	4 Stars High Review (4H)	1 star change
2B	two	5 Stars Low Review (5L)	4 Stars Low Review (4L)	1 star change
2C	two	4 Stars High Review (4H)	4 Stars Low Review (4L)	review level change
2D	two	5 Stars High Review (5H)	5 Stars Low Review (5L)	review level change

Table 4. Table of reputation conditions by *p5* star × review and *p4* star × review. The last column refers to the specific effect observed after taking the differences in investment amounts between *p5* and *p4* in each condition. These are the exact same conditions as that described in Table 2

effects. This would then gives us the exact average investment premium for a profile of having a review count of 11+ over having 1 extra star.

Figure 8 plots the differences in the mean differences of *p5* and *p4* investments for conditions 1C vs 1B, and 2C vs 2B. We see from the top plot that there is no significant premium between going from 4 stars to 5 stars versus going from 1-3 reviews to more than 10 reviews in world one (1C-1B) (mean difference of 0). However, this premium is quite large in world two ($\sim$ 5 extra credits). In world one, users value a 1 star change (from 4 stars to 5 stars) to be equivalent to having a review count increase of 10 or more, when differentiation in reputation between profiles is already low. However, there is a substantial difference in world two (2C-2B). From the bottom plot, which shows differences between 2C and 2B across intervals of *p5* review counts in 2C, we see that the overall premium increases with an increase in the number of additional reviews that a profile has.[7]. Substantively, this means that when there exists a profile that can reputationally differentiate itself from the others, it is much more beneficial for it to do so on reviews, rather than on stars. In fact, from the bottom plot of Figure 8 we can see at which exact review count threshold the review premium starts to kick in. There is a notable jump between 11-15 reviews and 16+, where the average investment difference goes from 0, to 10.

IMPLICATIONS FOR DESIGNERS

Our findings suggest that designing a reputation system that removes bias means maximizing the spread of reputations among the choices presented to users. This is in contrast to the common practice of simply presenting users with the best n-choices. We use Airbnb as an example. In Airbnb's current system, a potential user searching for an accommodation is always shown the top 10 listings with the highest reputation. Given this, it is unlikely that reputation plays a significant role in user choice since there is no clear differentiation of reputation signals among the alternatives. A more optimal system should thus show listings that have enough variation in their reputations as to make possible for the user to pick among a set of alternative accommodations that will more likely lead to his or her best choice.

Why will a platform want to implement such a system? Our findings show that number of reviews has a stronger effect than average rating, suggesting that longevity is privileged on the platform. This implies that users with more reviews will tend to attract more users. Developing a better system of differentiation of choices using the approach described above can then avoid systematically disadvantaging newer users, yet also ensure that biases displayed by users are kept in check. In such a case, we want to show a selection of listings to the user that would maximize the differentiation between the profiles yet minimize the potential of any one profile gaining significant advantage over the rest as a result of user bias created by say, the number of reviews that a particular listing may have. The byproduct of a more leveled playing field is a more efficient platform where supply is better able to keep up with demand. Our findings here provide some insights that designers consider when determining how to avoid a systematic disadvantage experienced by profiles due to user biases while still ensure that signals of quality can still be discerned. For example, the results in Figure 8 provides some clue as to at which interval a designer might want to "cut" when showing a group of profiles to users with different amounts of reviews in order to reduce the bias that profiles with less reviews may experience.

LIMITATIONS AND FUTURE WORK

Our current paper has a number of limitations, which can be further addressed in future work. First, we believe that further work is needed to understand exactly why participants responded the way they did in our study. Namely, we do not know for certain why subjects value review counts more than star ratings when assessing between two highly socially different profiles. Our hunch is that the higher review count serves as a better metric simply because it implicitly suggests that more people have vouched for the profile in the past. Second, due to sample restrictions, we could not conduct a more robust test of the threshold for the review premium that we found above. As such, the current study only serves as

[7]The wider confidence intervals at higher review counts is due to having smaller sample sizes of cases that has *p5* with higher review counts that increases the uncertainty around the mean

a rather coarse guide to show that such a threshold should exist on a given platform, with only preliminary evidence that any review difference of more than 10 would most likely trigger such a premium bias. We do not presume, however, that these thresholds are the same across platforms (in fact they most likely are not), and a more detailed examination of user behavior in response to them would prove to be beneficial.

CONCLUSION

We set out to accomplish two things with our current study. First, we wanted to test for the exact effects of the two core components of the conventional reputation system –average star ratings and review counts –on perceptions of trustworthiness. We tried to achieve this by conducting an investment game on a large population of Airbnb users. Second, we wanted to understand how design choices can affect assessment of trustworthiness by investigating users' responses to reputation contexts. We achieved this by constructing and randomly assigning subjects to two reputation conditions – "reputational contexts" –whereby we varied the distributional characteristics of the reputations of profiles we displayed to subjects.

We found that the effect of a profile going from having an average rating of 4 stars to 5 stars is roughly equivalent to having a review count increase of 10 or more –from having 1-3 reviews to having 11+ reviews. Yet we also found that when a reputation system has enough differentiating power, the number of review carries more weight than the average ratings. Indeed, what is interesting about these results is that the same exact reputation information can have varying efficacy for signaling trustworthy behavior depending on what alternatives that users have. Our work thus contributes to the existing work on reputation systems by quantifying exactly how two conventional feature of these systems affect users' perceptions of others' trustworthiness and identifying how their effects vary depending on various design choices.

ACKNOWLEDGEMENTS

We thank the anonymous reviewers for their detailed and helpful feedback. This work is supported by National Science Foundation Grant 1257138.

REFERENCES

1. Bruno Abrahao, Paolo Parigi, Alok Gupta, and Karen S Cook. 2017. Reputation offsets trust judgments based on social biases among Airbnb users. *Proceedings of the National Academy of Sciences* (2017), 201604234.
2. Joyce Berg, John Dickhaut, and Kevin McCabe. 1995. Trust, reciprocity, and social history. *Games and economic behavior* 10, 1 (1995), 122–142.
3. Abhijit Biswas and Edward A Blair. 1991. Contextual effects of reference prices in retail advertisements. *The Journal of Marketing* (1991), 1–12.
4. Peter M Blau. 1977. A macrosociological theory of social structure. *American journal of sociology* 83, 1 (1977), 26–54.
5. Gary Bolton, Ben Greiner, and Axel Ockenfels. 2013. Engineering trust: reciprocity in the production of reputation information. *Management Science* 59, 2 (2013), 265–285.
6. Gary E Bolton, Elena Katok, and Axel Ockenfels. 2004. How effective are electronic reputation mechanisms? An experimental investigation. *Management science* 50, 11 (2004), 1587–1602.
7. Chrysanthos Dellarocas. 2003. The digitization of word of mouth: Promise and challenges of online feedback mechanisms. *Management science* 49, 10 (2003), 1407–1424.
8. Chrysanthos Dellarocas. 2005. Reputation mechanism design in online trading environments with pure moral hazard. *Information Systems Research* 16, 2 (2005), 209–230.
9. Chrysanthos Dellarocas. 2006. How often should reputation mechanisms update a trader's reputation profile? *Information Systems Research* 17, 3 (2006), 271–285.
10. Andreas Diekmann, Ben Jann, Wojtek Przepiorka, and Stefan Wehrli. 2014. Reputation formation and the evolution of cooperation in anonymous online markets. *American Sociological Review* 79, 1 (2014), 65–85.
11. Tawanna R Dillahunt and Amelia R Malone. 2015. The promise of the sharing economy among disadvantaged communities. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 2285–2294.
12. Eyal Ert, Aliza Fleischer, and Nathan Magen. 2016. Trust and reputation in the sharing economy: The role of personal photos in Airbnb. *Tourism Management* 55 (2016), 62–73.
13. John Horton and Joseph Golden. 2015. Reputation inflation: Evidence from an online labor market. *Work. Pap., NYU* (2015).
14. Nan Hu, Jie Zhang, and Paul A Pavlou. 2009. Overcoming the J-shaped distribution of product reviews. *Commun. ACM* 52, 10 (2009), 144–147.
15. Daniel Kahneman. 2003. A perspective on judgment and choice: mapping bounded rationality. *American psychologist* 58, 9 (2003), 697.
16. Peter Kollock. 1999. The production of trust in online markets. *Advances in group processes* 16, 1 (1999), 99–123.
17. Debra Lauterbach, Hung Truong, Tanuj Shah, and Lada Adamic. 2009. Surfing a web of trust: Reputation and reciprocity on couchsurfing. com. In *Computational Science and Engineering, 2009. CSE'09. International Conference on*, Vol. 4. IEEE, 346–353.
18. Jeffrey A Livingston. 2005. How valuable is a good reputation? A sample selection model of internet auctions. *The Review of Economics and Statistics* 87, 3 (2005), 453–465.

19. Xiao Ma, Jeffrey T Hancock, Kenneth Lim Mingjie, and Mor Naaman. 2017. Self-Disclosure and Perceived Trustworthiness of Airbnb Host Profiles.. In *CSCW*. 2397–2409.

20. Mikhail I Melnik and James Alm. 2002. Does a sellerâĂŹs ecommerce reputation matter? Evidence from eBay auctions. *The journal of industrial economics* 50, 3 (2002), 337–349.

21. Paul Resnick, Ko Kuwabara, Richard Zeckhauser, and Eric Friedman. 2000. Reputation systems. *Commun. ACM* 43, 12 (2000), 45–48.

22. Paul Resnick and Richard Zeckhauser. 2002. Trust among strangers in Internet transactions: Empirical analysis of eBay's reputation system. In *The Economics of the Internet and E-commerce*. Emerald Group Publishing Limited, 127–157.

23. Paul Resnick, Richard Zeckhauser, John Swanson, and Kate Lockwood. 2006. The value of reputation on eBay: A controlled experiment. *Experimental economics* 9, 2 (2006), 79–101.

24. Juliet B Schor and Connor J Fitzmaurice. 2015. 26. Collaborating and connecting: the emergence of the sharing economy. *Handbook of research on sustainable consumption* 410 (2015).

25. Emily Sun, Ross McLachlan, and Mor Naaman. TAMIES: A Study and Model of Adoption in P2P Resource Sharing and Indirect Exchange Systems.

26. Arun Sundararajan. 2016. *The sharing economy: The end of employment and the rise of crowd-based capitalism*. Mit Press.

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

28. Georgios Zervas, Davide Proserpio, and John Byers. 2015. A first look at online reputation on Airbnb, where every stay is above average. (2015).