Keser, Claudia; Späth, Maximilian

Working Paper

The value of bad ratings: An experiment on the impact of distortions in reputation systems

cege Discussion Papers, No. 389

Provided in Cooperation with:
Georg August University of Göttingen, cege - Center for European, Governance and Economic Development Research

Suggested Citation: Keser, Claudia; Späth, Maximilian (2020) : The value of bad ratings: An experiment on the impact of distortions in reputation systems, cege Discussion Papers, No. 389, University of Göttingen, Center for European, Governance and Economic Development Research (cege), Göttingen

This Version is available at:
https://hdl.handle.net/10419/215743

Center for European, Governance and Economic Development Research
cege

Discussion Papers

Number 389 – April 2020

THE VALUE OF BAD RATINGS: AN EXPERIMENT ON THE IMPACT OF DISTORTIONS IN REPUTATION SYSTEMS

Claudia Keser
Maximilian Späth

GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN

ISSN: 1439-2305

The value of bad ratings: An experiment on the impact of distortions in reputation systems ♣

Claudia Keser[a], & Maximilian Späth[b]

April 5, 2020

Abstract

We study the robustness of reputation management systems against distortions in rating behavior. In a laboratory trust experiment with reputation management, we mimic a positive bias by exclusively offering the option to rate positively or to give no rating. As predicted by theoretical considerations, this bias leads to significantly less trust than a system that additionally offers a negative rating option. A system relying solely on negative ratings does not have such an adverse effect. This highlights the importance of negative ratings for the effectiveness of reputation systems.

Keywords: Trust, Trustworthiness, Reputation System, Experiment

JEL classification: C91, L14, C73

♣ We have received helpful comments from participants of the seminar on economics and management at the University of Paderborn, from participants of the CollEcons colloquium at the University of Goettingen, from participants of the 19th ZEW Summer Workshop for Young Economists, from participants of the annual conference of the Gesellschaft für experimentelle Wirtschaftsforschung (GfeW) in Kassel and from participants of the annual conference of the Verein für Socialpolitik (VfS) in Freiburg. We would like to thank Stephan Müller and two anonymous referees for their helpful comments and suggestions. *Declarations of interest*: none.

[a] *Corresponding author*: University of Goettingen, Platz der Göttinger Sieben 3, 37073 Göttingen, Germany, claudia.keser@uni-goettingen.de, and CIRANO, Montreal, Canada.

[b] University of Goettingen, Platz der Göttinger Sieben 3, 37073 Göttingen, Germany, maximilian.spaeth@uni-goettingen.de.

1. Introduction

Many (online-) markets like, for example, eBay, Marketplace at Amazon, Airbnb, or Uber rely on reputation systems allowing potential buyers of goods and services to be informed of the experiences that other customers have had with the respective seller. Reputation systems thus permit sellers to build a reputation of trustworthiness and gain trust with potential customers. Increased trust and trustworthiness in a market is likely to lead to more trade. From a theoretical perspective, reputation systems can be implemented to reduce inefficiencies that occur due to moral hazard in markets with asymmetric information (e.g., Bar-Isaac and Tadelis, 2008). Experimental research (e.g., Keser, 2003; Boero et al., 2009; Masclet and Pénard, 2012; Lumeau et al., 2015;) provides evidence for the power of reputation systems to enhance both investment and relative return in the "trust game" introduced by Berg et al. (1995). Still, in practice there are many open questions with respect to the design of efficient reputation systems (see Swamynathan et al., 2010, and Josang and Golbeck, 2009, for overviews).

One important issue is an apparent inflation toward favorable evaluations. Some unsatisfied buyers do not rate negatively, although they have the option to do so. This distortion toward positive ratings is a frequently observed phenomenon in (online)-reputation systems (see Tadelis, 2016, for a review).[1] Dellarocas and Wood (2008) and Bolton et al. (2013) find that some dissatisfied buyers do not rate negatively, when they must fear the seller's retaliation. Even without the possibility to directly retaliate, Nosko and Tadelis (2015) report a mismatch between the share of negative ratings and the relatively much higher share of complaints by buyers on eBay. Fradkin et al. (2017) find that 20 percent of Airbnb guests, who privately state that they would not recommend a host, still give a favorable public rating.

On some markets, a distortion toward positive ratings might be even exogenously caused by the design of the reputation systems. Li et al. (2016) describe that not providing a rating on the Chinese online-market Taobao is automatically interpreted as a positive evaluation. Social networks such as Facebook, Instagram and Twitter use "likes" or "hearts" as recommendations, but do not offer

[1] Likewise, in education, a grade inflation with a trend toward a higher share of good grades is apparent (e.g., Jewell at al., 2013).

a direct opportunity to show dislike.[2] In general, the giving of prizes and awards may be seen as the attribution of (mostly) positive evaluations.

The aim of our research is to analyze the effects that a positive (or negative) bias might have on trust and trustworthiness. We design a controlled laboratory experiment to measure if and how distortions built into the design of the rating system impact the effectiveness of the system to inform trustors and discipline trustees. Our experiment is based on the trust-and-reputation-management game introduced by Keser (2003). Trustors (which may represent customers of an online market) are invited to rate the trustees (sellers) after having interacted with them. In each of 20 periods, a trustor is randomly matched with a trustee under the constraint that a trustor never meets the same trustee more than once in a row. All the ratings that a trustee has received are made public to future trustors. Trustees are not informed about their reputation score.

In our baseline treatment POSNEG, participants can decide between a *positive*, a *negative*, or a *no rating* option. In treatment POS, we censor rating options to the *positive* and the *no rating* option. This mimics a positive bias that might, for example, be statistically induced by a high (nonmonetary) cost of giving a negative rating, or, by a system that by design is relying solely on positive evaluations. In our third treatment, NEG, rating options are censored to the *negative* and the *no rating* option. POS and NEG induce structural biases.

As predicted by our theoretical argumentation, we find that the structural positive bias in the reputation system leads to inefficiencies: trust is significantly lower in POS than in the baseline treatment POSNEG. It is also significantly lower than in NEG. We observe no significant difference in trust between NEG and POSNEG. Trustworthiness is significantly higher in NEG than in POS. The trustworthiness in POSNEG does not differ significantly from the trustworthiness in POS or NEG. Finally, the structural positive bias leads to lower payoffs for trustors and thus a greater inequality between market sides.

[2] In 2016, Facebook introduced the option to react with a sad or an angry emoji. In 2018 it was announced that a 'downvote' button would be tested. Neither of these changes introduces a dislike button that would directly oppose the "like" reaction.

2. Experimental design and hypotheses

The design of our computerized laboratory experiment is based on Keser (2003). Pairs of trustors and trustees interact in a repeated trust game (Berg et al., 1995) with random strangers matching (Andreoni, 1988). Participant roles do not alter during the experiment. In each of 20 rounds, trustors and trustees are endowed with 10 experimental currency units (ECU) each. They decide sequentially. In the first of three decision stages, trustors have the opportunity to send any share of their individual endowment (restricted to integer amounts between zero and the endowment) to the trustee they are matched with. By allowing for investments of zero, we give trustors the opportunity to refrain from interacting with trustees. Any positive amount sent is tripled by the experimenter. This represents a situation in which investments are beneficial from a societal perspective. In the second decision stage, the trustees may return any share of the amount received (restricted to integer amounts between zero and the tripled amount invested by the trustor). In the third stage, those trustors that have invested positive amounts in the first stage are requested to rate the trustworthiness of their trustee with respect to the amount returned. If no investment has been made, the system automatically records that *no rating* is given. A summary of all the received ratings will be visible to a trustee's future interaction partners, when they make their investment decisions. After every round, trustors and trustees are re-matched. The matching protocol satisfies the constraint that no pair is interacting more than once in a row, but is random otherwise. This implies that the strangers matching is imperfect since two players can meet again in a later period.[3]

With respect to the rating system we consider three treatments: POSNEG, POS and NEG. In our baseline treatment POSNEG, trustees can give a *positive*, a *negative* or *no rating*. This treatment is comparable to the system used by Keser (2003). A trustor is informed of the number of rounds in which the trustee received a *positive rating*, a *negative rating* or *no rating*, respectively. To enhance intuitive understanding, a *positive rating* is represented by a smiling face and a *negative rating* by a frowning face. The information on the number of *no ratings* does not differentiate between rounds where no investment was made and rounds where *no rating* was given. Taking into account the findings of Lumeau et al. (2015) and Boero et al. (2009) that trustees even react to ratings that are never communicated to other trustors or ratings that are communicated to trustors only after

[3] We acknowledge that we cannot entirely rule out that trustors use strategies trying to identify the trustee they are currently paired with.

they have made their investment decision, we attempt to isolate the pure reputation effect by not informing trustees about the ratings received.

In the POS treatment, we censor the reputation system to the options *positive rating* or *no ratin*g. Since we do not allow for *negative ratings* in this treatment, the *no rating* option will have to cover both the neutral and the negative experiences. Analogously, we censor the rating options to a *negative rating* or *no rating* in NEG. The *no rating* option will have to cover both the neutral and positive experiences. All participants, including the trustees, are informed of the reputation system that is applied.

Our experimental design warrants three comments. First, we refrain from allowing trustees to rate trustors, in order to prevent the distorting influence of a fear of retaliation. Second our treatment design could be interpreted as framing. It might be argued that the reputation systems in POS and NEG convey similar information under distinct frames. The frame itself might already impact behavior.[4] However, our treatment variation is more than a mere change in the frame. For the exhibited number of *no ratings* in POS, neutral evaluations are merged with automatically generated *no ratings* as well as with negative assessments. Similarly, neutral evaluations are merged with automatically generated *no ratings* as well as with positive assessments in NEG. Hence, participants in POS are not unequivocally informed about negative evaluations, while those in NEG are not unequivocally informed about positive evaluations. The exhibited number of *no ratings* provides little information by itself and cannot be interpreted as the direct counterpart to the number of *positive* or *negative ratings,* respectively. The third comment is that since in POS the *no rating* option will have to cover both the neutral and the negative experiences, we expect a lower threshold for giving a *positive rating* in POS than in POSNEG. This might create a positive statistical bias in ratings. Analogously, since in NEG the *no rating* option will have to cover both the neutral and the positive experiences, we expect a higher threshold for giving a *negative rating* in NEG than in POSNEG. This might create a negative statistical bias in ratings.

Analyzing our finitely repeated trust game with reputation management by backward induction, the subgame-perfect-equilibrium solution predicts no transactions and thus no ratings. Hence, variations of the reputation management system would make no difference. The trust game

[4] In a contract framework, for example, Imas, Sadoff, and Samek (2017) observe higher effort under loss than under gain contracts, whereas Quidt et al. (2017) find no significant framing effect.

represents a social dilemma: while individual rationality leads to zero investment by the trustor, collective rationality would require in each round the full investment of the trustor's endowment. Indeed, the experimental results of Berg et al. (1995) and others (see Johnson and Mislin, 2011, for a meta study) show that, even in one-shot games, most trustors do invest and many trustees return positive amounts. Bolton et al. (2005), for example, argue that such behavior might be due to limitations in people's ability to conduct backward induction. A number of experimental studies suggest, however, that the trustees' return transfers might be explained by other-regarding motives (e.g., Ashraf et al., 2006) and/or (intention-based) reciprocity (e.g., McCabe et al., 2003; Van den Bos et al., 2009). In other words, due to internalized social norms and values, trustees might derive more utility from reciprocating trust than from abusing trust. If we assume that with some probability a trustee is such a trustworthy type, trustors decide in a game with incomplete information. The trustors' decision to trust will, among others, depend on their willingness to assume social risks. Bohnet et al. (2004), for example, identify betrayal aversion in a binary-choice trust game.

Trust and trustworthiness can be sustained as an equilibrium outcome in infinitely repeated trust games with discounting (e.g., Kreps, 1990; Gibbons, 2001). Camerer and Weigelt (1988), Neral and Ochs (1992), Anderhub, Engelmann and Güth (2002), Brandts and Figueras (2003) as well as Grosskopf & Sarin (2010) theoretically and experimentally investigate finitely repeated binary-choice trust games with incomplete information. These studies consider reputation building equilibria similar to the models by Kreps and Wilson (1982) and Milgrom and Roberts (1982) for the chain-store game (Selten, 1978). The basic idea behind this approach is simple: if some trustees are intrinsically trustworthy, it might be profitable for untrustworthy trustees to build a reputation of being trustworthy at least until the final rounds of the game.

Similar reasoning applies, when we add a reputation system to the finitely repeated trust game among strangers. The reputation system offers trustees the opportunity to signal or at least pretend to be of a reciprocal and trust-honoring type. If trustors can be expected to consider these signals when making their investment decisions, having a good reputation has a strategic value to the trustee. Note that the modelling of this situation as a game with incomplete information is very complex, requires a number of strong assumptions and provides multiple equilibria. The

complexity of the model would be increased by the fact that we want to investigate different variants of reputation systems (POS, POSNEG and NEG).

Thus, let us take for granted the existence of reputation equilibria with positive investments by trustors and trustworthy behavior by trustees, at least until the final rounds of the game. Inspired by Ostrom (1998), we want to go from here with empirically grounded explanations in an attempt to build what Ostrom calls a second-generation model. Trustee's reputation, trust (trustor's investment) and trustworthiness (trustee's relative return) are all interconnected in a complex way in reputation equilibrium. Let us describe the core relationship between trustee's reputation, trust and trustworthiness as presented in Figure 1. These links are given by the sequential structure of the game. Moreover, consider that, theoretically, a good reputation may be considered as a signal of the trustee's trustworthiness. Indeed, many empirical and experimental studies have shown that trustors place more trust in a trustee with a good reputation (e.g., Resnick and Zeckhauser, 2002; Keser, 2003). In a theoretical model of repeated interaction in the trust game, the trustee's reputation depends on her/his trustworthiness in previous play. If, in the case of strangers' interaction, a reputation system is in place, experimental studies (e.g., Keser, 2003; Masclet and Pénard, 2012; Abraham et al., 2016) suggest that trustors use some kind of threshold strategy: to give a positive evaluation they require a specific minimum level of relative return. If this threshold is not reached, they require some lower threshold to give a neutral rating.

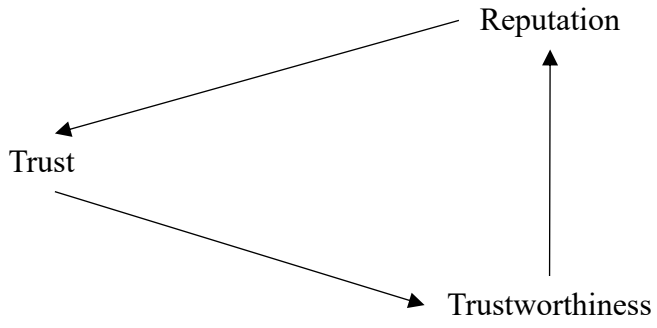

Figure 1: The core relationship between reputation, trust and trustworthiness.

Assume that, in the core relationship presented in Figure 1, the existence of a reputation management system has a strategic signaling value to the trustee. It directly affects the trustee's trustworthiness and, implicitly, the trustor's trust. Let us now consider how particular aspects of the reputation management system (as in POS, NEG and POSNEG) affect trustworthiness.

With respect to reputation giving, assume some threshold strategy as mentioned above. Trustors expect a minimum trustworthiness for a *positive rating*, as well as they expect a certain minimum in order not to give a *negative rating*. Thus, in POS there is a unique threshold for giving a p*ositive rating*, in NEG there is a unique threshold for not giving a *negative rating*, while in POSNEG the trustors decide based on two thresholds. Assume that the trustees anticipate this kind of behavior (without knowing the exact thresholds) and care for their ratings.

Comparing POS and POSNEG, the threshold strategy assumption implies that *positive rating* is one out of three potential reputation labels in POSNEG, where neutral and negative evaluations are strictly separated. A *positive rating* is probably more difficult to get in POSNEG than in POS (i.e., the threshold is higher in POSNEG than in POS), since in POS there exists only one other reputation label that comprises both neutral and negative evaluation. This relates to our argument given above for a positive statistical rating bias in POS. Consider now that, if a trustee receives a *positive rating* in an encounter it increases her/his overall reputation; if a trustee receives a *negative rating* in an encounter it decreases her/his reputation. In POSNEG, a trustee can receive both *positive ratings* and *negative ratings*. This means that if she/he receives a *positive rating* in one period, this positive signal to the trustors might be neutralized by a *negative rating* in a following period (for a better intuition, consider eBay's reputation score based on a "+1" for each *positive rating* and a "-1" for each *negative rating*). In contrast, in POS, once the trustee receives a *positive rating*, it cannot be neutralized any more: her/his overall reputation will remain positive until the end of the experiment.[5] For this reason, we might expect (at least until the final periods) more continual eagerness to receive a good rating in POSNEG than in POS. Together with the above threshold argument, this suggests overall higher trustworthiness in POSNEG than in POS.

In analogy to the difference between POS and POSNEG, we expect a higher threshold for not receiving a *negative rating* in NEG than in POSNEG. Additionally, consider that in NEG, once the

[5] Furthermore, recall that in our experiment the reception of no rating does not reveal whether the trustor did choose to give *no rating* or was not allowed to rate after having invested zero.

trustee receives a *negative rating*, it cannot be neutralized any more in that her/his overall reputation remains negative until the end of the experiment. Thus, we expect (1) that trustees might care almost as much about avoiding a *negative rating* in NEG as they care about receiving a *positive rating* in POSNEG, which (given the thresholds) reflects in their trustworthiness. At the same time, we expect (2) that in NEG the eagerness to avoid a *negative rating* is more persistent than the eagerness to receive a *positive rating* in POSNEG. Thus, overall, we do not expect significant differences in trustworthiness between POSNEG and NEG.

By transitivity, it follows that we should expect NEG to lead to more trustworthiness than POS. This is in keeping with empirical evidence by Shankar (2015), who finds that users at the online knowledge exchange Stack Overflow (a question-and-answers site for programmers) react more strongly, in terms of the quantity of contributions, to downvotes than to upvotes received to their answers. Similarly, Standifird (2001) and Lucking-Reiley et al. (2007) find a stronger impact of *negative ratings* than of *positive ratings* on prices on eBay. Finally, studies in psychology provide evidence for a fear of negative evaluation (Watson and Friend, 1969) and, in general, a stronger psychological effect of bad than of good events and information (Baumeister et al., 2001).

Reconsidering the core relationship presented in Figure 1, we assume that trustworthiness determines reputation. Reputation, in turn, determines trust. Therefore, we argue that trust shows the same pattern as trustworthiness. **Thus, we hypothesize to find higher trust and trustworthiness in POSNEG than in POS, similar trust and trustworthiness in POSNEG and NEG, and higher trust and trustworthiness in NEG than in POS.**

To test these hypotheses, we conducted our experiment in the period from 2016 to 2018 in the Göttingen Laboratory of Behavioral Economics at the University of Göttingen, Germany. Recruitment was done via ORSEE (Greiner, 2015). The experiment was programmed using z-Tree (Fischbacher, 2007). All instructions were read aloud by the same experimenter in all sessions. The instructions can be found in the Appendix. In total, 300 individuals participated in 21 sessions. For treatments POS and NEG, we collected the data of nine independent populations with ten participants (five trustors and five trustees) each. For POSNEG we collected data of twelve independent populations. Every session was concluded by a questionnaire. The average age of the

participants was around 24 and approximately 53 percent of them were female.[6] Each ECU earned during the experiment was converted to 0.03 euros. On average, participants stayed about 75 minutes in the laboratory and were paid around 14 euros, including a show-up fee of 4 euros.

3. Results

We denote the Wilcoxon rank-sum test as *rank-sum test* and the Wilcoxon matched-pairs signed-rank test as *signed-rank test*. Unless stated otherwise, we base the non-parametric tests on population averages, i.e., on nine or twelve observations per treatment. All tests are two-sided and we require $p = 0.05$ for significance.

3.1. Trust

Trust is measured by the investments of trustors. Comparing the amounts invested between treatments, we find evidence for an adverse impact of the structural positive bias. Table 1 conveys that average investments are lower in POS than in POSNEG. The decrease in trust of 14 percent is statistically significant (rank-sum test, $p = 0.033$). Furthermore, the average investments are significantly lower in POS than in NEG (rank-sum test, $p = 0.012$). The reputation system in NEG performs slightly better than the unrestricted system in POSNEG but the difference is statistically not significant (rank-sum test, $p = 0.749$). Hence, we find evidence our three hypotheses on trust.

Table 1: Descriptive statistics on investments per treatment.

Investment	Average	Median	Standard deviation
POSNEG	7.04	9	3.55
POS	6.07	7	3.89
NEG	7.11	10	3.68

Note: In ECU. Median and standard deviation are on the individual level.

[6] We find no significant differences in age (Kruskal-Wallis test, $p = 0.313$) and gender (Fisher's exact test, $p = 0.254$) between treatments.

Figure 2 conveys the average investments per period. In POSNEG and NEG investments are above those in POS in every period of the game, though the difference between POSNEG and POS in the very first period is only marginal. Indeed, considering the first period in isolation, we find no significant differences in investments between treatments.[7] This is in keeping with our assumption of an indirect impact of reputation management on trust. In the first half of the experiment, solely the difference between NEG and POS is statistically significant.[8] In the second half, both POSNEG and NEG show significantly higher average investments than POS.[9] For all treatments, Figure 2 suggests a typical endgame effect with relatively low investments in the last two periods. Comparing the last two periods with the average of the earlier periods, we find that the differences are not significant, though.[10]

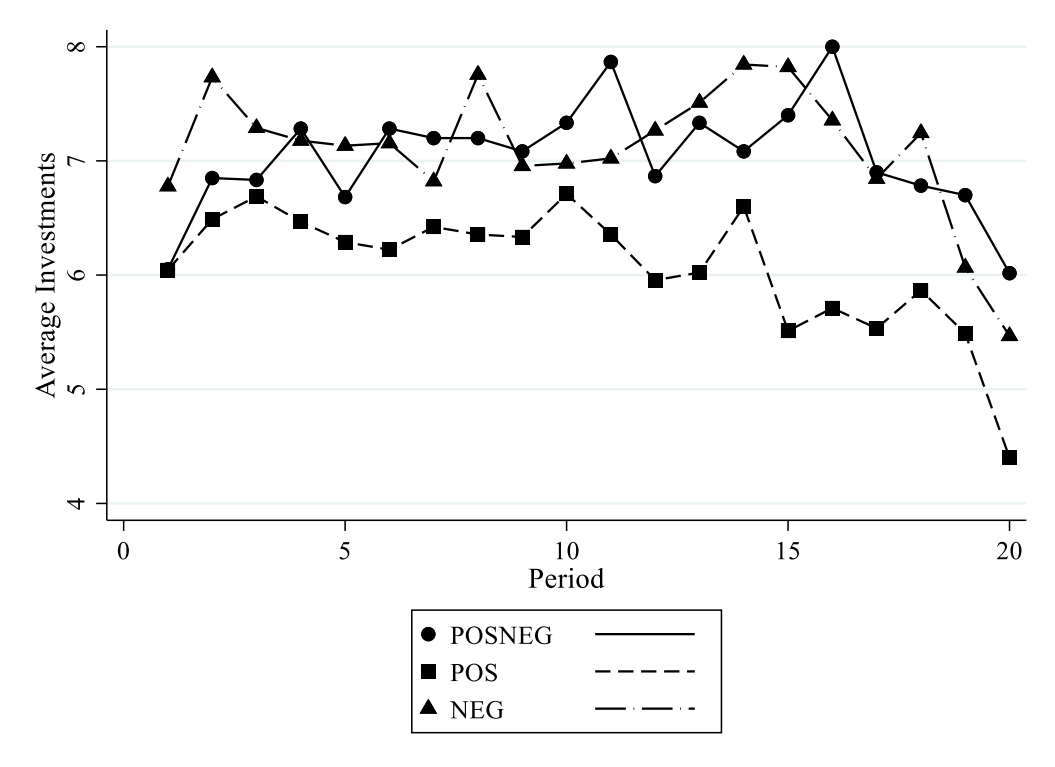

Figure 2: Average investments per period (by treatment).

[7] Individual level rank-sum tests: POSNEG vs POS, N = 105, p = 1.000; POSNEG vs NEG, N = 105, p = 0.297; POS vs NEG, N = 90, p = 0.377.
[8] Rank-sum-tests: POSNEG vs POS, p = 0.286; POSNEG vs NEG, p = 0.776; POS vs NEG, p = 0.042.
[9] Rank-sum-tests: POSNEG vs POS, p = 0.011; POSNEG vs NEG, p = 0.915; POS vs NEG, p = 0.004.
[10] Signed-rank tests: POSNEG, p = 0.136; POS, p = 0.051; NEG, p = 0.066.

Figure 3 provides the distribution of individual investment decisions. It reveals that more than half of the investments are either zero or 10 ECU. More concretely, it is the maximum investment of 10 ECU that is chosen most often. The relative frequencies are 0.55 in NEG, 0.48 in POSNEG and 0.44 in POS. The difference between NEG and POS is statistically significant, while the differences between POSNEG and POS and POSNEG and NEG are not.[11] The relative frequencies of zero investments are 0.10 in POSNEG, 0.11 in NEG, and 0.15 in POS. The differences between these shares are not statistically significant.[12] Note that the occurrence of zero investments might lead to an overestimation of the adverse effect of a positive bias on investments. The reason is that the reputation score that is presented to trustors does not differentiate between the *no ratings* that were actually given and *no ratings* that were automatically recorded. An automatically created *no rating* would be interpreted as a sign of untrustworthiness in POS, but as a sign of trustworthiness in NEG and as neutral information in POSNEG.

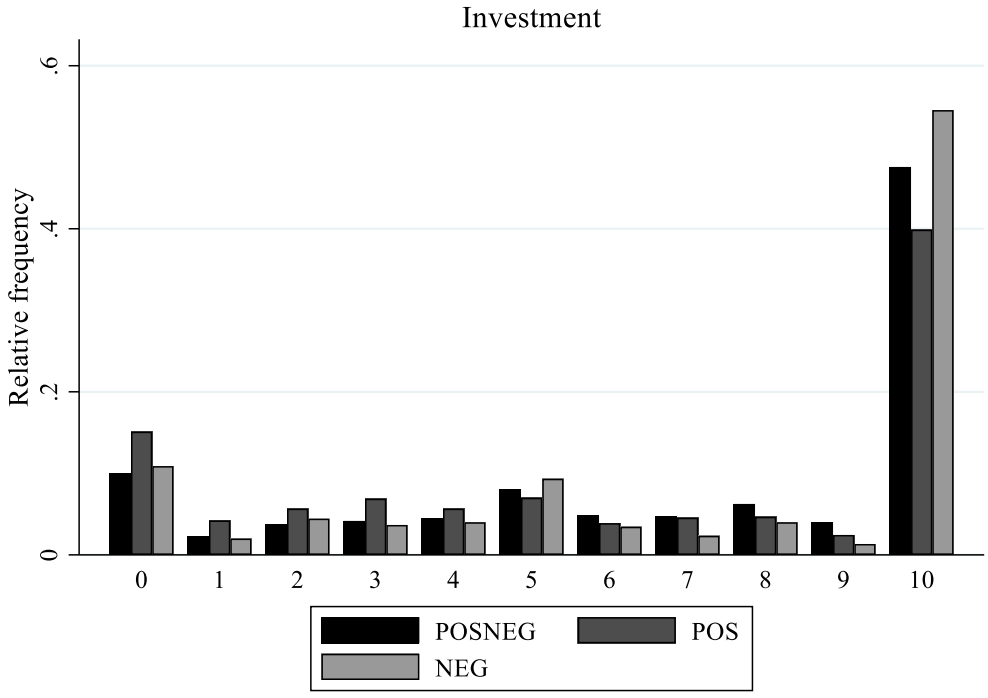

Figure 3: Relative frequency of individual investments (by treatment).

[11] Rank-sum-tests: POSNEG vs POS, p = 0.117; POSNEG vs NEG, p = 0.270; POS vs NEG, p = 0.031.
[12] Rank-sum-tests: POSNEG vs POS, p = 0.126; POSNEG vs NEG, p = 0.498; POS vs NEG, p = 0.251.

Interestingly, we find no significant correlation between the trustors' self-assessed degree of being risk-seeking and average investments on the population level.[13] The self-assessment was part of the ex-post questionnaire. Following Dohmen et al. (2011), we asked the question: "Generally speaking, are you a risk seeking person?". We used a 7-point Likert scale, with 1 = *not risk seeking at all* to 7 = *very risk seeking*. Our result is in keeping with what Houser et al. (2010) have found based on the Holt and Laury (2002) measure of risk attitudes.

3.2. Trustworthiness

Trustworthiness is measured by the relative returns of trustees. To calculate relative returns we divide the amount returned by the amount received. Table 2 reveals that, on average, we find the highest relative returns in NEG and the lowest in POS. Relative returns in POSNEG are in between and very close to those in NEG. Note that the level of trustworthiness that we measure is influenced by potentially untrustworthy trustees being taken out of the analysis since they do not receive an investment. However, as discussed in Section 3.1 on the distribution of investments, we do not observe significant differences in the shares of zero-investments across treatments and, therefore, no evidence of a distortion. We do not find a statistically significant difference in trustworthiness neither between POSNEG and POS (rank-sum test, $p = 0.155$) nor between POSNEG and NEG (rank-sum test, $p = 0.569$). However, we do find a significantly higher trustworthiness in NEG than in POS (rank-sum test, $p = 0.031$). Hence, we find evidence for two of the three hypotheses on trustworthiness.

Table 2: Descriptive statistics on relative returns per treatment.

Relative return	Average	Median	Standard deviation
POSNEG	0.51	0.50	0.197
POS	0.46	0.50	0.228
NEG	0.52	0.57	0.198

Note: Median and standard deviation are on the individual level.

[13] Spearman's rank correlation: $p = 0.124$.

Figure 4 shows the average relative return per period. It reveals that treatment effects, if they exist, are more pronounced in the first than in the second half of the experiment. In the first ten periods, average relative returns in NEG are above those in POSNEG in all but one and above those in POS in all rounds. Similarly, relative returns in POSNEG are above those in POS. Again, the difference between POS and NEG is statistically significant, while the other two differences are not.[14] In the second half of the game, we cannot detect any significant treatment effect.[15] An endgame effect with decreasing relative returns toward the end of the game is visible in all treatments. Comparing the last two periods—when the strategic value of reputation has vanished—with the average of the earlier periods, we find this endgame effect to be significant in POSNEG (signed-rank test, $p = 0.004$) and in NEG (signed-rank test, $p = 0.008$), but insignificant in POS (signed-rank test, $p = 0.173$). This might be seen as support of our theoretical consideration above that the strategic value of a favorable reputation (in the first 18 periods) is higher in POSNEG and in NEG than in POS, where any positive rating received can never be neutralized. Considering the first period in isolation, we find statistically significant differences in relative returns between POSNEG and POS (individual level rank-sum test, $N = 101$, $p = 0.033$) as well as NEG and POS (individual level rank-sum test, $N = 89$, $p = 0.012$), but not between POSNEG and NEG (individual level rank-sum test, $N = 102$, $p = 0.517$). This again, is in keeping with our theoretical assumption that reputation management impacts trustworthiness; trust is only indirectly affected. Recall that we found no significant difference in the trust exhibited in the first period. Additionally, in this period, we can detect no significant correlation between the investment of trustors and the relative returns of trustees on the individual level.[16] This suggests that the first-round differences in trustworthiness between treatments are not caused by the differences in trust, but by the fear of receiving a *negative rating* being stronger than the desire for a *positive rating.*

[14] Rank-sum tests: POSNEG vs POS, $p = 0.136$; POSNEG vs NEG, $p = 0.320$; POS vs NEG, $p = 0.015$.
[15] Rank-sum tests: POSNEG vs POS, $p = 0.155$; POSNEG vs NEG, $p = 0.887$; POS vs NEG, $p = 0.270$.
[16] Individual level Spearman's rank correlation: POSNEG, $N = 57$, $p = 0.500$; POS, $N = 44$, $p = 0.462$; NEG, $N = 45$, $p = 0.138$.

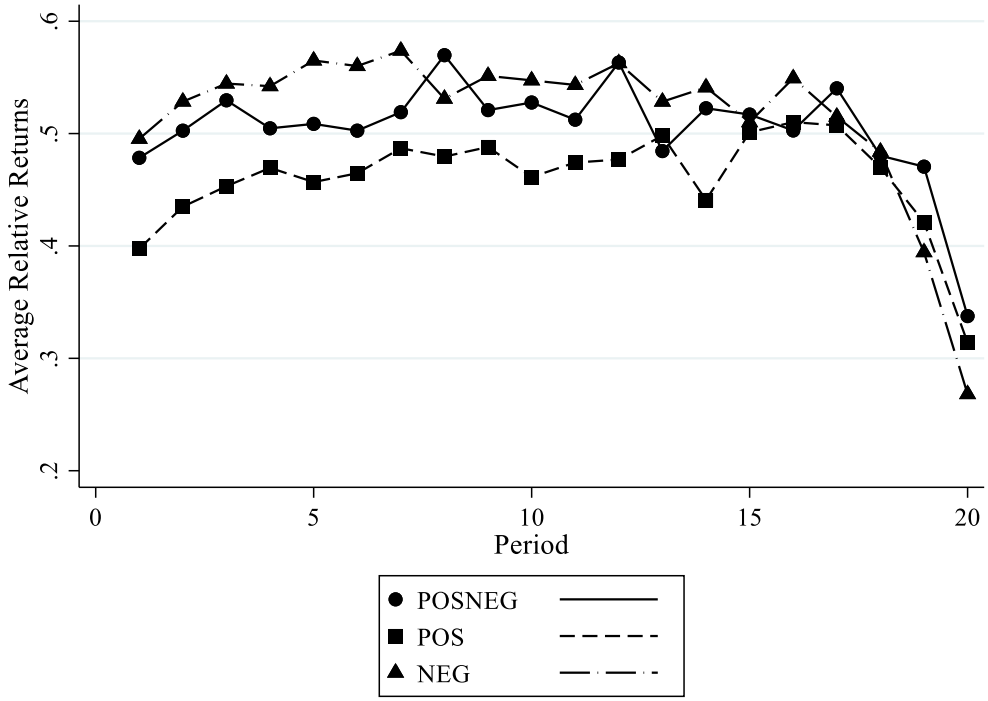

Figure 4: Average relative returns per period (by treatment).

Figure 5 displays the distribution of individual relative returns. We consider several prominent levels of relative returns and the intervals in-between. Concretely, we focus on relative returns of one, two thirds, one half, one third and zero. The most important level is a relative return of two thirds of the amount received. This share implies that trustor and trustee have identical payoffs in this period. The relative frequency of such a relative return is 0.24 in POS, 0.20 in POS, and 0.38 in NEG. The high share of trustees choosing this relative return in NEG suggests that trustees are eager to prevent a *negative rating*. On the population level, none of the differences between these relative frequencies are statistically significant, though.[17] Furthermore, we observe a bulk of relative returns at the equal split of the amount received. A further spike is at one third, the share to exactly return the trustor's investment. Finally, we consider a relative return of zero, which means a full exploitation of trust. Relative returns of more than two thirds, which would imply higher payoffs of trustors than of trustees, are very rare.

[17] Rank-sum tests: POSNEG vs POS, p = 0.500; POSNEG vs NEG, p = 0.088; POS vs NEG, p = 0.058.

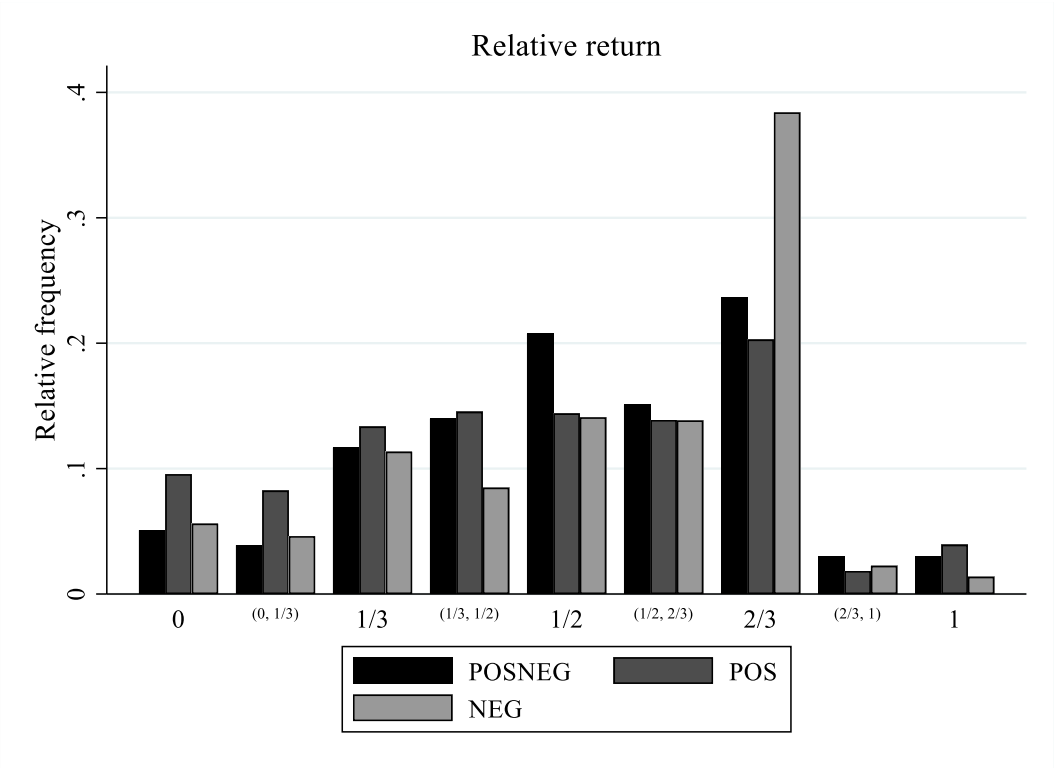

Figure 5: Relative frequency of individual relative returns (by treatment).

3.3. Ratings

Ratings are the connecting element between trustworthiness and trust. The cumulative distributions of relative returns per rating and treatment can be found in the Appendix (Figure 8). Comparing the distribution functions of *positive ratings* in POS and POSNEG as well as those of *negative ratings* in NEG and POSNEG, we do not observe important differences. We do observe, though, that the distribution functions of *no ratings* largely differ between POS, POSNEG and NEG. To provide some analysis on statistical significance, we consider that a *positive rating* corresponds to an average relative return of 0.60 (median 0.59) in POSNEG and to a relative return of 0.58 (median 0.57) in POS. The difference is not significant.[18] A *negative rating* corresponds to an average relative return of 0.29 (median 0.33) in POSNEG and of 0.33 (median 0.33) in NEG. The difference is not significant.[19] *No rating* in POSNEG corresponds to an average relative return of 0.44 (median 0.44). *No rating* in POS corresponds to an average relative return of 0.29 (median 0.33). Finally,

[18] Rank-sum test: p = 0.722.
[19] Rank-sum test: p = 0.434.

no rating in NEG corresponds to an average relative return of 0.62 (median 0.67). All differences are statistically significant.[20]

To summarize, only the distribution functions of *no rating* differ between treatments. This is due to the different meanings of *no rating*. In NEG, where *no rating* also comprises positive evaluations, a *no rating* is given for an average relative return of 0.62, which is about as high as the average relative return of 0.60 for which a *positive rating* is given in POSNEG. In POS, where *no rating* also comprises negative evaluations, a *no rating* is given for an average relative return of 0.29, which corresponds to the average relative return for which a *negative rating* is given in POSNEG.

Table 3: Average share of exhibited positive, nil, and negative ratings per treatment.

	Positive rating	No rating	Negative rating
POSNEG	0.624	0.182	0.193
POS	0.557	0.443	n. p.
NEG	n. p.	0.744	0.256

Note: It is not possible (n. p.) to receive *negative ratings* in POS or *positive ratings* in NEG.

Table 3 displays the shares of positive, nil, and negative ratings as exhibited to trustors. Recall, that these ratings include the automatically created *no ratings*. Considering our baseline treatment POSNEG, we observe a large share of *positive ratings* and smaller shares of *negative ratings* or *no ratings*. We observe a lower share of *positive ratings* in POS than in POSNEG. This difference is not statistically significant, though.[21] The lack of a *negative rating* option in POS apparently causes participants to give *no rating* instead. Similarly, in NEG, participants cannot give a *positive rating*; the best evaluation they may provide is a *no rating*. We observe a higher share of *negative ratings* in NEG than in POSNEG. The difference is statistically not significant, though.[22]

[20] Rank-sum test: POSNEG vs POS: p = 0.002; POSNEG vs POS: p = 0.001; POS vs NEG: p = 0.003.
[21] Rank-sum test: p = 0.177.
[22] Rank-sum test: p = 0.118.

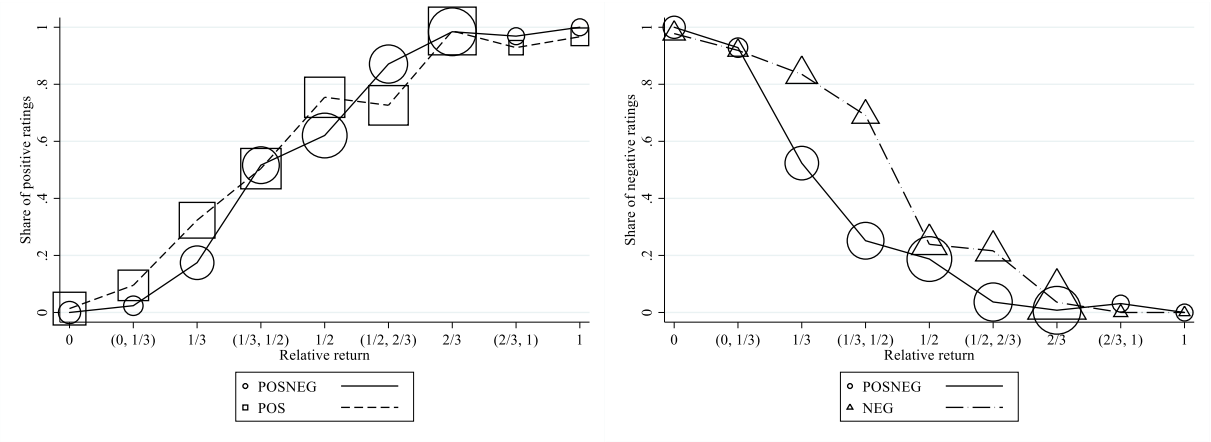

Figure 6: Left side: Share of *positive ratings* per level of relative returns (by treatment). Right side: Share of *negative ratings* per level of relative returns (by treatment). Note: The size of the circles, squares, and triangles increases with the number of cases with such a trustworthiness level in each treatment.

For a more detailed analysis of rating behavior, Figure 6 exhibits the share of *positive ratings* (left side) and *negative ratings* (right side) for each of the five levels of trustworthiness and the intervals in-between, as considered in Figure 5. The size of the circles, squares, and triangles increases with the number of cases with such a trustworthiness level in each treatment. We find that trustors tend to base their ratings on the relative returns of trustees. In POSNEG and POS, trustors give more often a *positive rating* the higher the relative return. In POSNEG and NEG, trustors give more often a *negative rating* the lower the relative return. Shares above 90 percent are reached, for *positive ratings*, at relative returns of two thirds and above in POSNEG and POS, and, for *negative ratings*, below relative returns of one third in POSNEG and NEG. In other words, the trustee can be 'pretty sure' to receive a *positive rating*, if she/he returns at least two thirds of the amount received. At the same time, the trustee can be 'pretty sure' to receive a *negative rating*, if she/he returns less than of one third. Note also that the trustee can be 'pretty sure' to avoid a *negative rating*, if she/he returns more than one half in POSNEG and at least two thirds in NEG.

Table 4: Multilevel mixed-effects logit regression on giving a *positive rating* (1) or *negative rating* (2)

	(1) *Positive rating*	(2) *Negative rating*
Relative return	21.474*** (1.236)	-21.283** (1.310)
POS	0.626 (0.458)	
NEG		1.709** (0.574)
constant	-9.619*** (0.655)	7.814*** (0.650)
level-3 variance	0.000	0.623
level-2 variance	4.413	4.009
N	1844	1882

Note: Standard errors in parentheses. Reference category for POS and for NEG: POSNEG. Column 1 regards treatments POSNEG and POS, Column 2 regards treatments POSNEG and NEG. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 4 provides the results of a multilevel mixed-effects logit regressions on rating giving of trustors. The use of a multilevel approach seems inevitable given that observations are clustered in subjects and populations. Table 4, Colum 1 displays the determinants of giving a *positive rating* in POSNEG or POS. We find a significantly positive coefficient for Relative return, indicating that trustors indeed base their rating on the trustworthiness of trustees in these two treatments. However, controlling for Relative return, we do not find differences in *positive rating* giving between POSNEG and POS. Similarly, we analyze the determinants of giving a *negative rating* in POSNEG or NEG (Table 4, Column 2). Again, we find a significantly negative coefficient for Relative return. In addition, the treatment dummy for NEG has a significantly positive coefficient. Since we control for Relative return, this indicates that trustors are more willing to give a *negative rating* in NEG than in POSNEG. This suggests that the threshold for <u>not</u> giving a *negative rating* is higher in NEG than in POSNEG. Recall that we do not observe such a difference for the giving of a *positive rating*, when we compare POSNEG and POS.

Turning to the reaction of trustors to reputation scores, Figure 7 displays the average investments per shares of *positive ratings* (left side) and of *negative ratings* (right side) in reputation scores. We arbitrarily choose intervals of 0.1 for the share of *positive*/*negative ratings*.[23] Again, the size of the circles, squares, and triangles increases with the number of cases with such a rating share in the treatment under consideration. We observe that investments are larger, the larger the share of *positive ratings*. Investments are smaller, the larger the share of *negative rating*s in the reputation scores. These observations indicate that ratings determine the investment level.

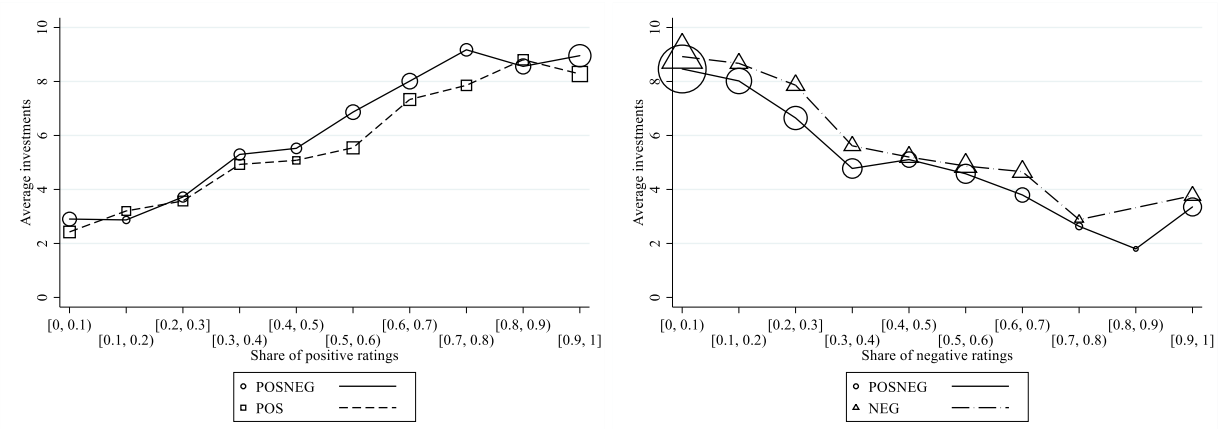

Figure 7: Left side: Share of *positive ratings* (in intervals of 0.1) and the corresponding average investment (by treatment). Right side: Share of *negative ratings* (in intervals of 0.1) and the corresponding average investment (by treatment).

We run multilevel mixed-effects regressions on the investment by trustors on the individual level. Table 5, Column (1) serves as a reference and confirms the adverse effect of a structural positive bias on investments: with POSNEG as the reference category, the coefficient of the treatment dummy POS is significantly negative. The coefficient of NEG is statistically not different from zero. Next, we add variables representing the reputation scores. Since, by design, reputation scores differ in their composition between treatments, we consider the comparisons of POSNEG and POS separately from the comparison of POSNEG and NEG. Column (2) in Table 5 displays the determinants of trustors' investments in POSNEG and POS. It shows that trustors react to the

[23] Note that not the shares but absolute numbers of ratings are visible to trustors. We use the intervals of shares for the purpose of illustration.

reputation score of the respective trustee they are interacting with: the coefficient of the Share of *positive ratings* in the reputation score is significantly positive. Controlling for the Share of *positive ratings,* we do not find treatment differences between POSNEG and POS. The statistically insignificant coefficient of the POS dummy in Column (2) suggests that the differences in trust between POSNEG and POS are not driven by differences in trustor's reactions to ratings between these treatments. The statistically insignificant interaction term in Column (3) confirms that trustors react to *positive ratings* similarly in POSNEG and POS. Table 5, Column (4) displays the determinants of investments in POSNEG and NEG. We find a significantly negative coefficient of the share of *negative ratings* but no significant treatment effect. Finally, the insignificant interaction term in Column (5) shows no differences in trustors' reaction to *negative ratings* between POSNEG and NEG.

To summarize, both the non-parametric analysis and the regression results suggest that trustors react to the share of *positive* and *negative ratings*, respectively. We find no differences of in this reaction, when we compare the behavior in POSNEG to the one in POS and NEG, respectively.

Table 5: Multilevel mixed-effects regression on investment

Investment	(1)	(2)	(3)	(4)	(5)
POS	-1.015** (0.379)	-0.586 (0.348)	-0.768 (0.414)		
Share of *positive ratings*		6.397*** (0.189)	6.261*** (0.252)		
Share of *positive ratings* x POS			0.311 (0.380)		
NEG	0.039 (0.379)			0.537 (0.424)	0.514 (0.440)
Share of *negative ratings*				-7.913*** (0.256)	-7.963*** (0.355)
Share of *negative ratings* x NEG					0.102 (0.512)
constant	7.089*** (0.248)	3.095*** (0.257)	3.179*** (0.277)	8.619*** (0.282)	8.629*** (0.286)
level-3 variance	0.000	0.000	0.000	0.265	0.265
level-2 variance	3.134	2.769	2.776	2.951	2.951
level-1 variance	10.567	6.575	6.571	6.612	6.612
N	2850	1995	1995	1995	1995

Note: Standard errors in parentheses. Only periods > 1 are considered. Reference category for POS and for NEG: POSNEG. Columns (2) and (3) regard treatments POSNEG and POS, Columns (4) and (5) regard treatments POSNEG and NEG. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

3.4. Payoffs

Considering the average payoff per round, we calculate that trustors (13.7 ECU) earn significantly less than trustees (19.8 ECU).[24] We find that the treatments have an impact on the payoffs of trustors: their payoffs in POS (12.7 ECU) are significantly lower than of those in POSNEG (14.1 ECU) and NEG (14.3 ECU).[25] A payoff greater than 10 ECU in a round shows us that the investment was profitable, in the sense that the amount returned exceeded the amount invested. We find a significantly lower share of profitable rounds in POS (0.58) than in POSNEG (0.71) or NEG (0.70).[26] The payoffs of trustees do not significantly differ between treatments.[27]

Finally, we analyze whether the trustors' payoff increases with trust. In POSNEG, higher average investments are significantly and positively correlated with higher payoffs (Spearman's rank correlation: rho = 0.867, p = 0.000). For POS (p = 0.088) and NEG (p = 0.188) the correlation fails significance. For trustees we detect no significant correlation between their average relative returns and their payoffs in POSNEG and POS, but a significantly negative correlation in NEG.[28]

4. Conclusion

Our theoretical considerations (including empirical and experimental evidence) as well as our experimental results demonstrate an adverse effect of a structural positive bias in reputation systems on trust: trustors' investments are significantly lower in POS than in POSNEG or NEG. The system in NEG performs as well as the unrestricted system in POSNEG in that investments reach approximately the same level in both treatments. The analysis regarding the trustworthiness of trustees displays a similar picture in that the relative returns are significantly lower in POS than in NEG. The relative returns in POSNEG are between those in the two other treatments, but they are not significantly different from either of them. Thus, we do not find an effect on trust of a structural negative bias in reputation systems. The treatment effects on trust and trustworthiness translate into differences in payoffs between treatments. The positive bias in POS leads, among the

[24] Rank-sum test: p = 0.000.
[25] Rank-sum tests: POSNEG vs POS, p = 0.039; POSNEG vs NEG, p = 0.972; POS vs NEG, p = 0.015.
[26] Fisher's exact test: POSNEG vs POS, p = 0.000; POSNEG vs NEG, p = 0.467; POS vs NEG, p = 0.000.
[27] Rank-sum tests: POSNEG vs POS, p = 0.477; POSNEG vs NEG, p = 0.887; POS vs NEG, p = 0.423.
[28] Spearman's rank correlation: POSNEG: p = 0.146; POS: p = 0.380; NEG: rho = -0.833, p = 0.002).

three treatments, to the lowest payoffs of trustors and thus to the largest inequality between market sides.

Fisher et al. (2018) suggest that we can apply our results also to five-star rating systems. They find evidence that customers think in categories of positive ratings (four or five stars) and negative ratings (one or two stars). We do not intend to advocate the restriction of rating options, when there is no need to do so. The negative reputation system has some important downsides. Trustees have only a limited possibility to reconcile their reputation score after a *negative rating*. Trustors do not receive information that might be relevant for them. Unfortunately, our experimental design does not allow any inference on how this restricted information transmission is affecting beliefs. Nevertheless, our analysis highlights the high value of *negative ratings*.

There are several potential ways to combat a positive bias and to motivate customers to truthfully give a negative rating, when they are unsatisfied. As described by Bolton et al. (2013) and Klein et al. (2016) it is of importance that costumers need not fear a seller's retaliation. Furthermore, not providing a rating should be made visible as no rating or a neutral rating. In order to impede fake rating only costumers that have been involved in a transaction should be allowed to evaluate (Mayzlin et al., 2013) and reviews might be sorted with respect to usefulness by automatic software (as, for example, used by YELP). Costumers should be able to easily report any attempt of sellers to prevent a negative rating. It might be considered to reduce the identifiability of raters, since a possible identification seems to promote the transmission of positive but not of negative signals (Rockenbach and Sadrieh, 2012). Negative ratings might be framed in a more positive way. By officially linking the best rating (e.g., five of five stars) to a "normal" quality, any intermediate rating is a criticism framed in a more positive way. Finally, we want to highlight the importance of designing a closed market in the sense that participants do not have the possibility to leave the market and return with new identities. Yamagishi and Matsuda (2002) conduct a market experiment in which the true quality of a seller's product is private information. They find that in an open market a positive reputation system with a range of ratings from 0 (neutral) to very good (+2) leads to more trustworthy behavior of sellers than a negative system with ratings from -2 (very negative) to 0 (neutral). This is due to the observation that sellers with highly negative reputation scores tend to rejoin the market with a new identity. The results of Yamagishi and Matsuda (2002) emphasize

the importance of solving the re-entry problem through measures like demanding an entry fee or using cryptographic identifiers, as suggested by Friedman and Resnick (2001).

References

Abraham, Martin, Veronika Grimm, Christina Neeß, and Michael Seebauer. 2016. "Reputation Formation in Economic Transactions." *Journal of Economic Behavior & Organization* 121:1–14. https://doi.org/10.1016/j.jebo.2015.10.010.

Anderhub, Vital, Dirk Engelmann, and Werner Güth. 2002. "An Experimental Study of the Repeated Trust Game with Incomplete Information." *Journal of Economic Behavior & Organization* 48 (2): 197–216. https://doi.org/10.1016/S0167-2681(01)00216-5.

Andreoni, James. 1988. "Why free ride?" *Journal of Public Economics* 37 (3): 291–304. https://doi.org/10.1016/0047-2727(88)90043-6.

Ashraf, Nava, Iris Bohnet, and Nikita Piankov. 2006. "Decomposing Trust and Trustworthiness." *Experimental Economics* 9 (3): 193–208. https://doi.org/10.1007/s10683-006-9122-4.

Bar-Isaac, Heski, and Steven Tadelis. 2008. *Seller Reputation.* Hanover, MA: Now Publishers.

Baumeister, Roy F., Ellen Bratslavsky, Catrin Finkenauer, and Kathleen D. Vohs. 2001. "Bad Is Stronger Than Good." *Review of General Psychology* 5 (4): 323–70. https://doi.org/10.1037//1089-2680.5.4.323.

Berg, Joyce, John Dickhaut, and Kevin McCabe. 1995. "Trust, Reciprocity, and Social History." *Games and Economic Behavior* 10 (1): 122–42. https://doi.org/10.1006/game.1995.1027.

Boero, Riccardo, Giangiacomo Bravo, Marco Castellani, and Flaminio Squazzoni. 2009. "Reputational Cues in Repeated Trust Games." *The Journal of Socio-Economics* 38 (6): 871–77. https://doi.org/10.1016/j.socec.2009.05.004.

Bohnet, Iris, and Richard Zeckhauser. 2004. "Trust, Risk and Betrayal." *Journal of Economic Behavior & Organization* 55 (4): 467–84. https://doi.org/10.1016/j.jebo.2003.11.004.

Bolton, Gary, Ben Greiner, and Axel Ockenfels. 2013. "Engineering Trust: Reciprocity in the Production of Reputation Information." *Management Science* 59 (2): 265–85. https://doi.org/10.1287/mnsc.1120.1609.

Bolton, Gary E., Elena Katok, and Axel Ockenfels. 2004. "How Effective Are Electronic Reputation Mechanisms? An Experimental Investigation." *Management Science* 50 (11): 1587–1602. https://doi.org/10.1287/mnsc.1030.0199.

Bolton, Gary E., Elena Katok, and Axel Ockenfels. 2005. "Cooperation Among Strangers with Limited Information About Reputation." *Journal of Public Economics* 89 (8): 1457–68. https://doi.org/10.1016/j.jpubeco.2004.03.008.

Brandts, Jordi, and Neus Figueras. 2003. "An Exploration of Reputation Formation in Experimental Games." *Journal of Economic Behavior & Organization* 50 (1): 89–115. https://doi.org/10.1016/S0167-2681(02)00042-2.

Camerer, Colin, and Keith Weigelt. 1988. "Experimental Tests of a Sequential Equilibrium Reputation Model." *Econometrica* 56 (1): 1. https://doi.org/10.2307/1911840.

Dellarocas, Chrysanthos, and Charles A. Wood. 2008. "The Sound of Silence in Online Feedback: Estimating Trading Risks in the Presence of Reporting Bias." *Management Science* 54 (3): 460–76. https://doi.org/10.1287/mnsc.1070.0747.

Dohmen, Thomas, Armin Falk, David Huffman, Uwe Sunde, Jürgen Schupp, and Gert G. Wagner. 2011. "Individual Risk Attitudes: Measurement, Determiants, and Behavioral Consequences." *Journal of the European Economic Association* 9 (3): 522–50. https://doi.org/10.1111/j.1542-4774.2011.01015.x.

Fischbacher, Urs. 2007. "z-Tree: Zurich toolbox for ready-made economic experiments." *Experimental Economics* 10 (2): 171–78. https://doi.org/10.1007/s10683-006-9159-4.

Fisher, Matthew, George E. Newman, Ravi Dhar, Gita Johar, and Stijn van Osselaer. 2018. "Seeing Stars: How the Binary Bias Distorts the Interpretation of Customer Ratings." *Journal of Consumer Research* 31 (2): 191. https://doi.org/10.1093/jcr/ucy017.

Fradkin, Andrey, Elena Grewal, and David Holtz. 2017. "The Determinants of Online Review Informativeness: Evidence from Field Experiments on Airbnb." MIT Sloan School of Management and Airbnb Working Paper.

Friedman, Eric J., and Paul Resnick. 2001. "The Social Cost of Cheap Pseudonyms." *Journal of Economics & Management Strategy* 10 (2): 173–99. https://doi.org/10.1111/j.1430-9134.2001.00173.x.

Gibbons, R. (2001). Trust in social structures: Hobbes and Coase meet repeated games. In K. S. Cook (ed.), *Russell Sage foundation series on trust* 2: 332–353. Russell Sage Foundation.

Greiner, Ben. 2015. "Subject pool recruitment procedures: Organizing experiments with ORSEE." *Journal of the Economic Science Association* 1 (1): 114–25. https://doi.org/10.1007/s40881-015-0004-4.

Grosskopf, Brit, and Rajiv Sarin. 2010. "Is Reputation Good or Bad? An Experiment." *American Economic Review* 100 (5): 2187–2204. https://doi.org/10.1257/aer.100.5.2187.

Holt, Charles A., and Susan K. Laury. 2002. "Risk Aversion and Incentive Effects." *American Economic Review* 92 (5): 1644–55. https://doi.org/10.1257/000282802762024700.

Houser, Daniel, Daniel Schunk, and Joachim Winter. 2010. "Distinguishing Trust from Risk: An Anatomy of the Investment Game." *Journal of Economic Behavior & Organization* 74 (1-2): 72–81. https://doi.org/10.1016/j.jebo.2010.01.002.

Imas, Alex, Sally Sadoff, and Anya Samek. 2017. "Do People Anticipate Loss Aversion?" *Management Science* 63 (5): 1271–84. https://doi.org/10.1287/mnsc.2015.2402.

Jewell, R. Todd, Michael A. McPherson, and Margie A. Tieslau. 2013. "Whose fault is it? Assigning blame for grade inflation in higher education." *Applied Economics* 45 (9): 1185–1200. https://doi.org/10.1080/00036846.2011.621884.

Johnson, Noel D., and Alexandra A. Mislin. 2011. "Trust Games: A Meta-Analysis." *Journal of Economic Psychology* 32 (5): 865–89. https://doi.org/10.1016/j.joep.2011.05.007.

Josang, Audun, and Jennifer Golbeck (2009). "Challenges for Robust Trust and Reputation Systems". 5th International Workshop on Security and Trust Management.

Keser, Claudia. 2003. "Experimental games for the design of reputation management systems." *IBM Systems Journal* 42 (3): 498–506. https://doi.org/10.1147/sj.423.0498.

Klein, Tobias J., Christian Lambertz, and Konrad O. Stahl. 2016. "Market Transparency, Adverse Selection, and Moral Hazard." *Journal of Political Economy* 124 (6): 1677–1713. https://doi.org/10.1086/688875.

Kreps, David. 1990. "Corporate Culture and Economic Theory'', in James Alt and Karl Shepsle (eds.), *Perspectives on Positive Political Economy*. New York: Cambridge University Press, 90–143.

Kreps, David M., and Robert Wilson. 1982. "Reputation and imperfect information." *Journal of Economic Theory* 27 (2): 253–79. https://doi.org/10.1016/0022-0531(82)90030-8.

Li, Lingfang, Steven Tadelis, and Xiaolan Zhou. 2016. *Buying Reputation as a Signal of Quality: Evidence from an Online Marketplace.* Cambridge, MA: National Bureau of Economic Research.

Lucking-Reiley, David, Doug Bryan, Naghi Prasad, and Daniel Reeves. 2007. "Pennies from eBay: The Determinants of Price in Online Auctions." *Journal of Industrial Economics* 55 (2): 223–33. https://doi.org/10.1111/j.1467-6451.2007.00309.x

Lumeau, Marianne, David Masclet, and Thierry Pénard. 2015. "Reputation and social (dis)approval in feedback mechanisms: An experimental study." *Journal of Economic Behavior & Organization* 112: 127–40. https://doi.org/10.1016/j.jebo.2015.02.002.

Masclet, David, and Thierry Pénard. 2012. "Do Reputation Feedback Systems Really Improve Trust Among Anonymous Traders? An Experimental Study." *Applied Economics* 44 (35): 4553–73. https://doi.org/10.1080/00036846.2011.591740.

Mayzlin, Dina, Yaniv Dover, and Judith Chevalier. 2014. "Promotional Reviews: An Empirical Investigation of Online Review Manipulation." *American Economic Review* 104 (8): 2421–55. https://doi.org/10.1257/aer.104.8.2421.

McCabe, Kevin A., Mary L. Rigdon, and Vernon L. Smith. 2003. "Positive Reciprocity and Intentions in Trust Games." *Journal of Economic Behavior & Organization* 52 (2): 267–75. https://doi.org/10.1016/S0167-2681(03)00003-9.

Milgrom, Paul, and John Roberts. 1982. "Predation, Reputation, and Entry Deterrence." *Journal of Economic Theory* 27 (2): 280–312. https://doi.org/10.1016/0022-0531(82)90031-X.

Neral, John, and Jack Ochs. 1992. “The Sequential Equilibrium Theory of Reputation Building: A Further Test.” *Econometrica* 60 (5): 1151. https://doi.org/10.2307/2951542.

Nosko, Chris, and Steven Tadelis. 2015. “The Limits of Reputation in Platform Markets: An Empirical Analysis and Field Experiment.” Cambridge, MA: National Bureau of Economic Research.

Ostrom, Elinor. 1998. “A Behavioral Approach to the Rational Choice Theory of Collective Action: Presidential Address, American Political Science Association, 1997.” *American Political Science Review* 92 (1): 1–22. https://doi.org/10.2307/2585925.

Quidt, Jonathan de, Francesco Fallucchi, Felix Kölle, Daniele Nosenzo, and Simone Quercia. 2017. “Bonus versus penalty: How robust are the effects of contract framing?” *Journal of the Economic Science Association* 3 (2): 174–82. https://doi.org/10.1007/s40881-017-0039-9.

Resnick, P. and Zeckhauser, R. (2002), "Trust among strangers in internet transactions: Empirical analysis of eBay' s reputation system", Baye, M. (ed.) *The Economics of the Internet and E-commerce* (*Advances in Applied Microeconomics, Vol. 11*) 127-156. Emerald Group Publishing Limited, Bingley. https://doi.org/10.1016/S0278-0984(02)11030-3.

Rockenbach, Bettina, and Abdolkarim Sadrieh. 2012. “Sharing information.” *Journal of Economic Behavior & Organization* 81 (2): 689–98. https://doi.org/10.1016/j.jebo.2011.10.009.

Selten, Reinhard. 1978. “The Chain Store Paradox.” *Theory and Decision* 9 (2): 127–59. https://doi.org/10.1007/BF00131770.

Shankar, Ramesh. 2015. Online Reputational Loss Aversion: Empirical Evidence from StackOverflow.com. Working paper. http://dx.doi.org/10.2139/ssrn.2603303.

Standifird, S. 2001. “Reputation and e-commerce: eBay auctions and the asymmetrical impact of positive and negative ratings.” *Journal of Management* 27 (3): 279–95. https://doi.org/10.1016/S0149-2063(01)00092-7.

Swamynathan, Gayatri, Kevin C. Almeroth, and Ben Y. Zhao. 2010. “The Design of a Reliable Reputation System.” *Electronic Commerce Research* 10 (3-4): 239–70. https://doi.org/10.1007/s10660-010-9064-y.

Tadelis, Steven. 2016. “Reputation and Feedback Systems in Online Platform Markets.” *Annual Review of Economics* 8 (1): 321–40. https://doi.org/10.1146/annurev-economics-080315-015325.

van den Bos, Wouter, Eric van Dijk, Michiel Westenberg, Rombouts, Serge A R B, and Eveline A. Crone. 2009. “What Motivates Repayment? Neural Correlates of Reciprocity in the Trust Game.” *Social cognitive and affective neuroscience* 4 (3): 294–304. https://doi.org/10.1093/scan/nsp009.

Watson, D., and R. Friend. 1969. “Measurement of Social-Evaluative Anxiety.” *Journal of consulting and clinical psychology* 33 (4): 448–57. https://doi.org/10.1037/h0027806.

Yamagishi, Toshio, and Masafumi Matsuda. 2002. “Improving the lemons market with a reputation system: An experimental study of internet auctioning.” University of Hokkaido, Technical Report.

Appendix

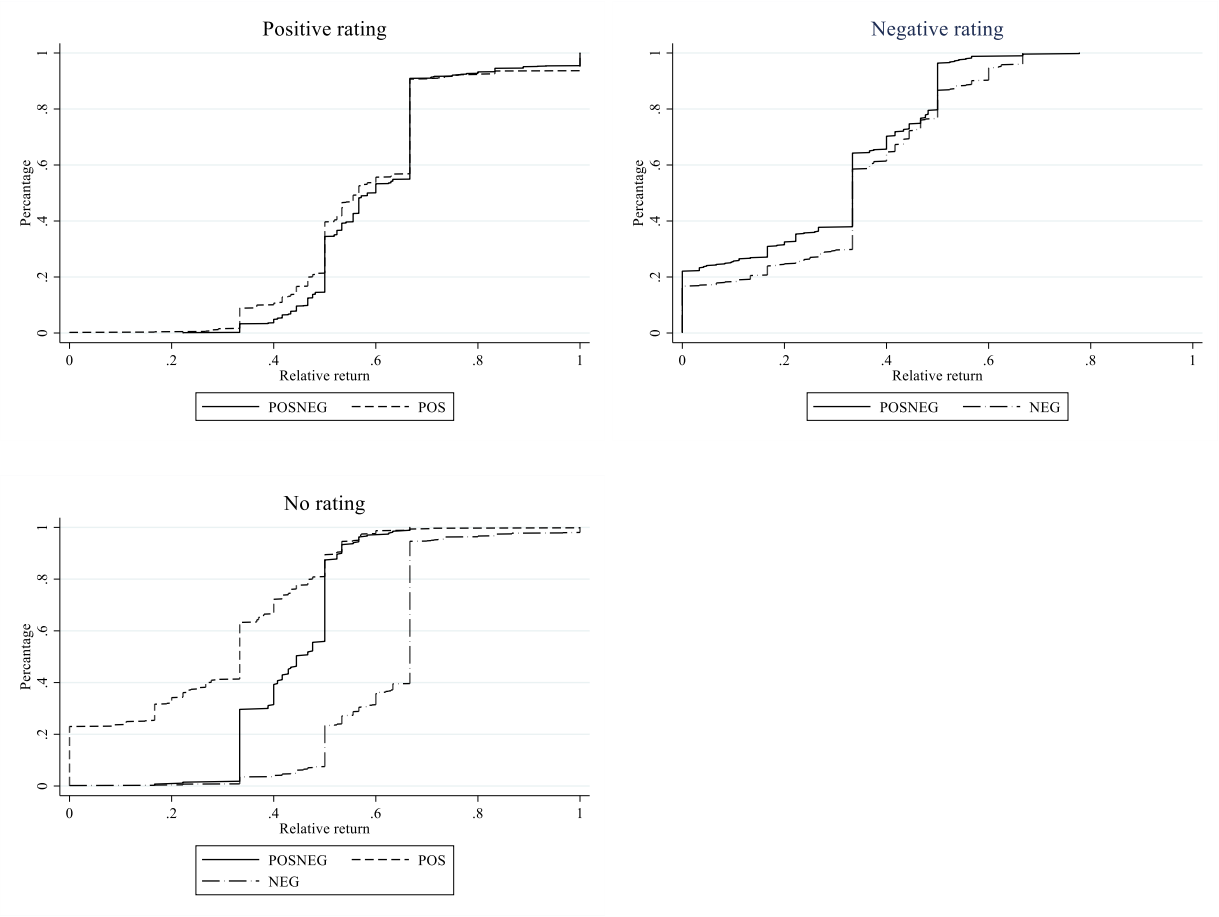

Figure 8: Cumulative distribution of relative returns corresponding to a *positive rating* (upper left), a *negative rating* (upper right) and *no rating* (lower left), per treatment.