Krügel, Jan Philipp; Paetzel, Fabian

Conference Paper

The Impact of Fake Reviews on Reputation Systems and Efficiency

Beiträge zur Jahrestagung des Vereins für Socialpolitik 2021: Climate Economics

Provided in Cooperation with:
Verein für Socialpolitik / German Economic Association

Suggested Citation: Krügel, Jan Philipp; Paetzel, Fabian (2021) : The Impact of Fake Reviews on Reputation Systems and Efficiency, Beiträge zur Jahrestagung des Vereins für Socialpolitik 2021: Climate Economics, ZBW - Leibniz Information Centre for Economics, Kiel, Hamburg

This Version is available at:
https://hdl.handle.net/10419/242415

The Impact of Fake Reviews on Reputation Systems and Efficiency*

Jan Philipp Krügel† and Fabian Paetzel‡

Sunday 28th February, 2021

Abstract

Online interactions are frequently governed by reputation systems that allow users to evaluate each other after an interaction. Effective reputation systems can increase trust and may improve efficiency in market settings. In recent years, however, fake reviews have become increasingly prevalent. Since it is difficult to clearly identify fake reviews in field studies, we design a laboratory experiment. Using a repeated public good game with a reputation system, we study (i) how feedback manipulation influences the reliability of average ratings and (ii) whether the existence of manipulated ratings reduces efficiency. We find that feedback manipulation generally decreases the reliability of average ratings in comparison to a control treatment where cheating is not possible. When manipulation is possible and free, average ratings become less reliable, expectations are lower and both cooperation and efficiency are significantly reduced. When there are costs of manipulation, however, average ratings are more reliable and contributions and efficiency are not impaired. Interestingly, this is the case even when costs of manipulation are comparatively low.

Keywords: Reputation Systems; Fake Reviews; Reliability of Ratings; Efficiency

JEL: C72; C91; D83; L14

*We thank Matthias Greiff, Ben Greiner, Timo Heinrich, Rupert Sausgruber and Stefan Traub as well as the participants of the economics workshop at the Helmut Schmidt University, the Vienna Business and Economics University and the 2019 ESA World Meeting in Vancouver. The research was carried out at WISO laboratory at the University of Hamburg in 2018 and 2019.

†Corresponding author. Helmut Schmidt University, Department of Economics and Social Science and Research Group FOR2104 (sponsored by the German Research Foundation), Holstenhofweg 85, 22043 Hamburg, Germany, Phone +49-40-6541-2432, kruegel@hsu-hh.de.

‡Helmut Schmidt University, Department of Economics and Social Science, Holstenhofweg 85, 22043 Hamburg, Germany, Phone +49-40-6541-3192, fpaetzel@hsu-hh.de.

1 Introduction

Many online platforms use reputation systems that allow users to evaluate each other after an interaction. Websites using reputation systems not only include businesses such as eBay, Amazon, TripAdvisor and Yelp, but also online communities such as Stack Exchange. A number of studies have shown that effective reputation systems increase trust among parties involved in an interaction (e.g. Ba & Pavlou 2002). In online marketplaces, reputation systems can reduce risks for customers by lowering information asymmetries and can help realize efficiency-gains for market participants (Hui et al. 2016, Tadelis 2016). Furthermore, the existence of a reputation system can have positive effects on trust, expectations, cooperation and efficiency in experimental settings such as trust games (Masclet & Pénard 2012), prisoner's dilemma games (Stahl 2013) and public good games (Greiff & Paetzel 2016, 2020).[1]

In order for a reputation system to be effective, it is vital that the users of the service can rely on the feedback provided by others. Reliability is high when the quality of the product or the past behavior of the interaction partner is strongly correlated with the rating displayed by the online platform. However, studies indicate that online reputation systems might suffer from several problems that negatively affect the reliability of overall (i.e., average) ratings. The first concern is selection bias. In particular, customers who are either extremely satisfied or dissatisfied are more likely to submit a review than those who have had a moderate experience (Hu et al. 2017, Karaman 2020, Schoenmueller et al. 2020). Another problem arises when feedback is assigned sequentially. In these situations, the first-mover might not report a bad experience by assigning a negative rating in order to avoid receiving a negative rating in return from the second-mover (Bolton et al. 2013, Fradkin et al. 2018, Bolton et al. 2019, Masclet & Pénard 2012).

Furthermore, and given its prevalence perhaps most importantly, there is evidence that many users of online platforms post or buy manipulated reviews of themselves or their products in order to artificially improve their reputation. Scientific studies (Ott et al. 2012, Mayzlin et al. 2014, Luca & Zervas 2016) and recent media reports indicate that the problem of fake reviews is widespread, although the precise extent of the problem is difficult to estimate.[2] To tackle the problem, online platforms such as Yelp and Amazon use algorithms and filters to identify and remove content that is thought to be fraudulent. Algorithms commonly use IP addresses, language or content in order to detect fake reviews, although many firms do not disclose much about the methodology or the extent to which potential fake reviews are removed.[3] If a fake review is detected, the review is deleted and users may receive fines or be banned from using the service.[4] In order to avoid detection,

[1] For surveys of the literature, see Chen et al. (2020), Greiff & Paetzel (2020), Athey & Luca (2019), Luca (2017) and Dellarocas (2003).

[2] While Ott et al. (2012) suspect that up to 6% of reviews on sites like Yelp and TripAdvisor may be deceptive, Luca & Zervas (2016) report that 16 % of reviews on Yelp are identified by Yelp's algorithm as potential fakes. The consumer group Which Travel, which is based in the United Kingdom, analyzed a total of almost 250,000 reviews of popular tourist destinations. According to their research, one in seven of the hotels had "blatant hallmarks" of fake reviews, while others raised "serious concerns". TripAdvisor has reacted to the reports and has removed these reviews (The Guardian 2019).

[3] In a sector inquiry, the German Federal Cartel Office (*Bundeskartellamt*) found that online platforms use a variety of different measures to identify fake content (German Federal Cartel Office 2020). Some websites put restrictions on those who can post reviews: on Amazon.com, for example, a review can only be submitted if a user has spent at least $50 in the past 12 months.

[4] In a recent settlement, the American Federal Trade Commission fined the two companies "Devumi" and "Sunday

some businesses are willing to spend money on fake reviews that are not easily distinguished from real reviews. Several websites openly sell fake reviews on the Internet. The independent German consumer organization Stiftung Warentest recently reported that fake reviews are available for 10 Euros per review (Stiftung Warentest 2020).

This raises the following question: How does the existence of manipulated feedback influence (i) the reliability of average ratings, and (ii) trust in average ratings and the efficiency of the reputation system? In addition, algorithms and other detection devices may mitigate the negative effects of fake reviews on overall review quality as they introduce costs of manipulation. Businesses continuously relying on fake reviews face the prospect of getting fined for manipulation or they have to buy "better" fake reviews that are not easily detected as fakes and hence removed. Therefore, a second important question is whether introducing manipulation costs has a positive influence on (i) and (ii) compared to a scenario without manipulation costs.

In field studies, it is difficult to clearly identify if a review is fake or real. This is acknowledged by several authors (e.g. Mayzlin et al. 2014, Luca & Zervas 2016). As a result, the extent of fake reviews on a platform can only be estimated and answering the aforementioned questions with field data is naturally difficult. This study therefore addresses these questions using a laboratory experiment. In contrast to field studies, we are always aware whether a rating has been manipulated or not and we can precisely estimate the reliability of average ratings.

Our experiment builds on Greiff & Paetzel (2016) and Greiff & Paetzel (2020). The participants play a repeated public good game with a group size of two and absolute stranger matching to imitate simplified market interactions. Using a public good game allows the players to easily rate how much the other person has contributed to a collective outcome.[5] At the end of each period, participants evaluate each other by assigning a rating, and, in the next stage, receive information on how they were evaluated by their partner. At this juncture, participants can manipulate (i.e., improve) the rating that the partner has given them and thus try to artificially improve their reputation. We conduct three treatments in which the cost to improve the rating varies: no costs, low costs or high costs. Then, at the beginning of the next period, participants receive information about their own and their new partner's average rating of the preceding periods, but they are not informed on whether the rating of the new partner has been manipulated or not. As a control, we conduct an additional treatment where manipulation is not possible.

The setting allows us to study the impact of manipulated feedback on several variables, including individual evaluation behavior, reliability of ratings, trust (positive expectations) and overall efficiency of the reputation system. We can not only identify fake ratings unambiguously but also exclude other potential problems of reputation systems, including the negative effects of selection bias and sequential feedback: In our experiment, participants cannot refrain from rating others, interact only once and evaluate each other simultaneously. One may argue that in reality, users

Riley Skincare" for faking reviews of their products, which was deemed deceptive marketing (Federal Trade Commission 2019). In a similar case, 19 companies who were writing or commissioning fake reviews had to pay more than $350,000 in penalties (New York State Office of the Attorney General 2013). In Canada, the national Competition Bureau fined the telecommunication provider Bell $1.25 million for encouraging employees to write fake online reviews for their phone apps (CBC News 2015).

[5]One could also use a trust game or a market experiment to explore our research questions. We chose a public good game because it allows us to compute a clean measure for the quality of the ratings that are displayed to the subjects. Furthermore, the overriding mechanism driving reputation and reliability of ratings should be independent of the underlying game and therefore be applicable to other circumstances.

of online platforms usually cannot improve the ratings received by others, which is possible in our design, but rather improve their rating by posting additional positive fake reviews. However, notice that the outcome of manipulated ratings and positive fake reviews is the same: The average rating of a subject increases in both cases. Hence, we believe that our experimental approach offers insights into effects of fake reviews on reputation systems that can complement field studies on the subject. To the best of our knowledge, this study is the first experiment that enables participants to manipulate feedback given by others.

We find that many subjects manipulate their rating if given the chance. Hence, the ability to manipulate feedback decreases the reliability of average ratings. In particular, average ratings are less informative about past behavior when subjects are able to manipulate the rating received by their transaction partner without incurring any costs. Consequently, subjects do not trust average ratings in this scenario and contributions and overall efficiency are significantly lower than in the other treatments. In the two treatments with manipulation costs, on the other hand, there is less cheating and average ratings are more reliable. Interestingly, while the reliability of ratings is higher with high manipulation costs than with low manipulation costs, contributions and overall efficiency are on similar levels in both treatments. Our findings suggest that reputation systems will not work effectively if an online platform does not remove fake content or does not punish the manipulation of reviews. Reputation systems are more efficient when there are costs of manipulation. However, the manipulation costs do not necessarily have to be high: For the reputation system to work effectively, comparatively low costs of manipulation are sufficient. Our results support anecdotal evidence about fake reviews: Even though most users are aware that fake reviews exist, they continue to consult and partially rely on online feedback when making their decisions, especially when the platform has a system in place that at least removes the most obvious fake reviews.

Our paper not only contributes to research on reputation systems, but also speaks to the prominent literature about cheating behavior in economic experiments (e.g. Fischbacher & Föllmi-Heusi 2013, Charness et al. 2014, Gneezy et al. 2018, Abeler et al. 2019, Necker & Paetzel 2020). This branch of literature stresses that (i) subjects cheat, (ii) but not to the full extent, and (iii) subjects' cheating behavior does not necessarily depend on the costs and benefits of a lie. Abeler et al. (2019) find evidence that subjects have a preference for being honest and a preference for being seen as honest. Their conclusions are based on a meta-study analyzing behavior in the very specific "roll-a-die and report the outcome" experiment, which was first developed by Fischbacher & Föllmi-Heusi (2013). Gneezy et al. (2018) argue that lying aversion explains why subjects do not lie to the full extent. Our study contributes to this literature by analyzing cheating behavior in a fundamentally different experiment with repeated interactions. In the presence of reputation systems in repeated interactions, the incentive to build a reputation to gain from cooperation might dominate a preference for being (perceived as) honest. In line with previous findings on cheating behavior in other experiments, our results also indicate that subjects cheat, albeit not to the full extent. However, as opposed to previous literature, we find clear evidence that the net outcome of a lie (a function of costs and expected benefits) significantly matters for the decision to cheat. In addition, we find that in the treatment with no cheating costs, the preference for "having a good reputation" seems to be more important than a preference for being honest.

The organization of this paper is as follows. In Section 2 we present the experimental design and treatments and motivate the main research question in further detail. Results are presented in Section 3 while Section 4 concludes the paper.

2 The Experiment

2.1 Experimental Design

The experimental design is based on previous experimental work on reputation systems by Greiff & Paetzel (2015, 2016, 2020). In this study, we add a manipulation stage to the experiment. The participants in our experiment play a repeated public good game over 15 periods with varying partners. In each period, participants are randomly and anonymously paired and each participant makes several decisions.

First, we elicit the participants' expectations about their partners' contributions. Second, the subjects simultaneously choose how much of their endowment ($e = 3$) they want to contribute to a public good ($c \in \{0, 1, 2, 3\}$). Assume participant i is being matched with participant j. Then, i's initial payoff is given by $\pi_i(c_i, c_j) = 4(e_i - c_i) + 3(c_i + c_j) - 2$. Third, after participants are informed about choices (contributions) and initial payoffs, each participant evaluates the contribution decision of their partner. Participants simultaneously assess each other's decision by assigning between 0 and 10 stars. The participants are explicitly told that 0 stars corresponds to the worst and 10 stars to the best possible rating.[6] Fourth, participants are informed about the rating that the partner has given them. At this juncture, our treatment variation comes into place (see Table 1). In the treatment *No costs*, subjects can improve the rating that the partner has given them without incurring any costs.[7] In the treatments *Low costs* and *High costs*, subjects are able to buy additional stars, with one star costing 20 cents in *Low costs* and 60 cents in *High costs*. If a subject buys additional "fake" stars in these treatments, then the costs are subtracted from the initial payoff described above. The real rating of the partner and the fake stars that a subject adds to her evaluation make up the rating for a round in *No costs*, *Low costs* and *High costs*. In the *Control* treatment, the manipulation of the rating given by the partner is not possible and subjects go straight to the next period.

Treatment	Manipulation	Costs of adding an additional "fake" star
Control	Not possible	–
No costs	Possible	0 cents
Low costs	Possible	20 cents
High costs	Possible	60 cents

Table 1: Treatment Structure

Afterwards, participants are re-matched. In the next period, participants receive information about their new partner's average rating and their own average rating of the three preceding periods. The displayed average rating in the treatments *No costs*, *Low costs* and *High costs* may include round ratings that have been manipulated. However, subjects are not informed on whether the average rating of the new partner has in any way been manipulated in the preceding periods or not.

Henceforth, we will use the terms "real rating" for the rating received by the partner (which consists of "real stars"), "fake stars" for additional stars added to the real rating in each period,

[6] Obviously, participants can still assess whether 10 stars is indeed the "best" rating possible from their point of view.

[7] Reducing the rating is not possible in any of the treatments.

"round rating" for the sum of real stars and fake stars in each period and "displayed rating" for the average round rating that is displayed to the subjects.[8]

2.2 Conjectures

Generally, a reputation system in a public good game enables two mechanisms that may foster cooperation. First, subjects may cooperate with those who have earned a good reputation (*indirect reciprocity*). Second, subjects may cooperate with those whom they expect to cooperate as well, and use reputation as a predictor of cooperation (*conditional cooperation*). As shown by Greiff & Paetzel (2016), the introduction of a reputation system that is based on evaluations increases contributions to the public good compared to baseline game with no reputation mechanism, largely because ratings are indeed informative about the past behavior of subjects: Despite the fact that ratings are subjective and may contain noise, there is a positive correlation between cooperation and ratings, and this is common knowledge. Importantly, cooperation only increases when participants know not only the rating of their partner, but also their own rating. Using information about their own ratings, subjects are able to form second-order beliefs. Positive first- and second-order beliefs are a necessary precondition for conditional cooperation. The use of average ratings in our design is based on Greiff & Paetzel (2020), whose results indicate that contributions are highest when subjects receive information about their own average rating and the average rating of their partner. Displaying average ratings also allows us to study the impact of fake reviews on the reliability of overall ratings: As mentioned in the introduction, positive fake reviews and manipulated ratings both increase the average rating of a subject.

The first question we want to answer is whether the ability to manipulate feedback decreases the reliability of displayed ratings and if this leads to a less efficient outcome. For this purpose, we compare the treatments *Control* and *No costs*. The results of our *Control* treatment are taken from Greiff & Paetzel (2020). Hence, we know that in *Control*, displayed ratings are informative about past behavior and lead to higher contributions compared to a scenario without a reputation system. In *No costs*, on the other hand, the subjects have no way of knowing whether the displayed rating of their partner is real or manipulated. Furthermore, adding additional fake stars to the real rating has no consequences for the participants in this treatment on any level. Hence, we hypothesize that displayed ratings are *not* informative about past behavior in *No costs*. If this is the case, displayed ratings are useless and we expect contributions to be lower than in *Control* since the two mechanisms that foster cooperation (indirect reciprocity and conditional cooperation) are not as easily available to subjects as with truthful evaluations.

Main Hypothesis *In contrast to the Control treatment, displayed ratings in the treatment No costs are not informative about the past behavior of participants. We expect contributions and overall efficiency to be lower in No costs than in Control.*

Our second research question addresses whether introducing manipulation costs in the treatments *Low costs* and *High costs* has a positive influence on the reliability of displayed ratings

[8] In period 2, displayed ratings are round ratings from period 1. In period 3, displayed ratings are averages over round ratings in periods 1 and 2. Starting in period 4, displayed ratings are computed based on the last three round ratings.

compared to the *No costs* treatment. If there are manipulation costs, subjects may add fewer stars to their real rating than in *No costs* or may even refrain from buying fake stars altogether. If this is the case, displayed ratings should be more informative about past behavior in *Low costs* and *High costs* than in *No costs*. As a consequence, contributions and overall efficiency may also be higher in *Low costs* and *High costs* than in *No costs*.

We conjecture that subjects add fewer fake stars to their rating compared to *No costs* when the manipulation costs are sufficiently high. However, we consider it an empirical question how high the costs have to be in our experiment in order to change participants' behavior in the hypothesized direction. We therefore conduct two treatments with differing costs. In *High costs*, adding an additional fake star to the real rating costs 60 cents. Hence, subjects are always able to buy ten additional stars and still have a positive payoff in every possible scenario (1 euro in the worst-case scenario; compare the payoff table in Figure 8, Appendix A). Choosing 20 cents in *Low costs* is to some extent arbitrary. We choose this amount because (i) it is considerably less than 60 cents and (ii) still has a noticeable impact on participants' payoff.

2.3 Procedures

For the three treatments in which manipulation was possible we ran four sessions with 18 participants each, and in the control treatment we ran five sessions with 18 participants each.[9] In total, 306 participants took part in our experiment, which resulted in 4590 observations over 15 periods of play. The sessions lasted for about 80 minutes. We programmed the experiment using z-tree (Fischbacher 2007). Participants were mostly undergraduate students from various disciplines, recruited via the software hroot (Bock et al. 2014). The experiment was carried out at WISO laboratory at the University of Hamburg.

All participants received a show-up fee of 5 euros. In addition, they received a variable payoff consisting of two parts. One period was randomly selected and participants' contributions in this period determined the first part of the variable payoff. If a subject was in one of the treatments *Low costs* or *High costs* and bought fake stars in this period, the costs were subtracted from the payoff. Then, another period was randomly selected. If a participant's expectation in this particular period was correct, the second part of the variable payoff was 4 euros. If the expectation was incorrect, the second part of the variable payoff was 0 euros. Since two different periods were randomly selected, hedging was not possible. All this was common knowledge among participants. Average payments (including the show-up fee) were 18.05 euros.

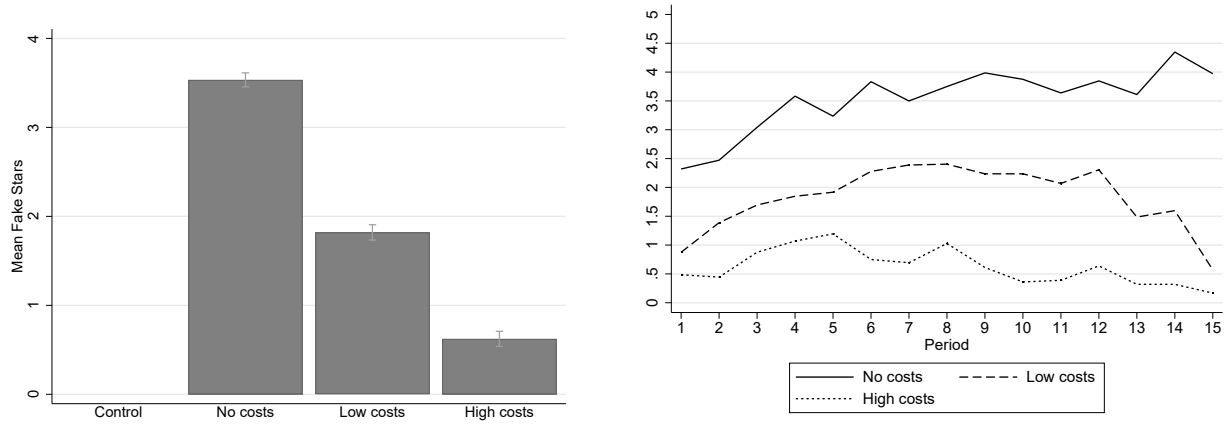
3 Results

The analysis of the impact of fake stars on reputation systems and efficiency makes it necessary to dissect the complete transmission channel from evaluation behavior to cheating behavior, the forming of expectations and contribution behavior. In Subsection 3.1 we dissect (i) how and to what extent participants make use of fake stars, (ii) how participants evaluate each other and (iii) how reliable round ratings and displayed ratings are in each treatment. In Subsection 3.2 we analyze contribution behavior and subsequently efficiency. In Subsection 3.3 we analyze expectations in

[9] Observations of the treatment *Control* are borrowed from the treatment AVE3 in Greiff & Paetzel (2020).

further detail. The analysis of expectations dissects how participants decide on their contributions based on both displayed ratings and opportunities to cheat.

3.1 Fake Stars, Evaluation Behavior and Reliability of Displayed Ratings



(a) Average fake stars by treatment pooled over periods.

(b) Fake stars over time; data points are means of the number of fake stars for each period.

Figure 1: Subfigure (a) shows average numbers of fake stars. Subfigure (b) presents fake stars over time.

Figure 1 shows to what extent participants make use of fake stars. The left panel (Figure 1a) shows that with no costs, cheating is almost twice as high as in the case of low costs and about four times higher than in case of high costs. It can be taken from this bar chart that, with increasing costs, average fake stars decrease substantially and significantly.[10] In the treatment *No costs*, only 3 out of 72 subjects never added a single fake star. In the treatment *Low costs*, 18 of 72 subjects never bought a fake star and in *High costs*, 31 of 72 subjects never bought a fake star. Or to put it the other way around, the fraction of cheaters is decreasing with increasing costs for fake stars.

The right panel (Figure 1b) shows that the treatment differences persist continuously over time. Interestingly, in *No costs*, fake stars seem to increase over time, whereas in *High costs*, average fake stars seem to persist on a low but constant level. Figure 1b also shows an end-game effect in the treatments *Low costs* and *High costs*, whereas no end-game effect occurs in the treatment *No costs*.

The descriptive analysis is confirmed by random effects regressions, displayed in Table 2, in which we regress fake stars on the partner's contribution (c_j), the participant's own contribution (c_i), the participant's own - (e_i) and the partner's recently received real rating (e_j), period (1,...,15) and some interactions. We include indicator variables for treatments, taking *High costs* as the baseline category. *Control* is not integrated because adding fake stars was not possible in this treatment.

Regressions (1) - (6) in Table 2 show that both treatment dummies are positive and significant, meaning that in *No costs* and in *Low costs*, the usage of fake stars is significantly higher than in the baseline *High costs*. Additionally, a Wald-test shows that both treatment dummies are significantly

[10] Non parametric tests corroborate this relationship with high significance ($p < 0.001$).

	(1)	(2)	(3)	(4)	(5)	(6)
No costs	2.911***	1.789***	3.427***	2.100***	3.050***	1.853***
	(0.259)	(0.371)	(0.257)	(0.325)	(0.265)	(0.375)
Low costs	1.196***	0.885*	1.579***	1.433***	1.278***	0.935*
	(0.286)	(0.394)	(0.291)	(0.372)	(0.286)	(0.394)
c_j			0.240***	0.266***		
			(0.065)	(0.064)		
c_i			-0.002	0.022		
			(0.070)	(0.070)		
e_j			-0.483***	-0.490***		
			(0.030)	(0.030)		
e_i			0.026	0.016		
			(0.021)	(0.020)		
disappointment					0.538**	0.566***
(c_i/(1+e_j))					(0.167)	(0.168)
relative evaluation					0.424**	0.450**
(c_j/(1+e_i))					(0.152)	(0.150)
Period	0.021	-0.038**	-0.030*	-0.091***	0.034*	-0.030*
	(0.015)	(0.014)	(0.013)	(0.017)	(0.015)	(0.014)
No costs×Period		0.140***		0.170***		0.151***
		(0.034)		(0.028)		(0.035)
Low costs×Period		0.039		0.019		0.043
		(0.030)		(0.027)		(0.030)
Constant	0.453*	0.931***	2.559***	3.052***	0.058	0.553**
	(0.181)	(0.202)	(0.296)	(0.334)	(0.193)	(0.205)
N	3240	3240	3240	3240	3240	3240
$R^2(within)$	0.001	0.010	0.433	0.447	0.014	0.024
$R^2(between)$	0.307	0.307	0.442	0.440	0.292	0.291
$R^2(overall)$	0.118	0.123	0.434	0.443	0.120	0.126

Table 2: Fake stars as a function of partner's contribution c_j, participant's own contribution c_i, treatment dummies *No costs* and *Low costs*. e_i is the participants' own real rating assigned to the partner and e_j is the partner's real rating. *High costs* is the baseline treatment. Regressions are random effects regressions with robust standard errors. Observations from *Control* are not considered. Significance levels: $*$ for $p < 0.05$, $**$ for $p < 0.01$, $***$ for $p < 0.001$.

different. Unsurprisingly, the usage of fake stars is highest when it is free (*No costs*). The costs of fake stars are taken into account systematically by participants. One might interpret these findings as evidence of a rational decision by participants to use fake stars.

Regression (2) in Table 2 depicts that the interaction variable *No costs*×Period is positive and significant. This confirms the visual inspection above. There is a positive time trend only in *No costs*. Regressions (4) - (6) analyze possible motivations for making use of fake stars in further detail. It turns out that the use of fake stars increases with the partner's contribution (c_j) and decreases with the real rating (e_j). Another way to understand fake behavior can be achieved through analyzing 'disappointment' as the relation between the participant's own contribution and

the real rating received from the interaction partner (c_i/(1+e_j)), considering the participant's own relative real rating compared to the partner's contribution ((c_j/(1+e_i)). Regressions (5) and (6) show two things: (i) The higher the disappointment, the more fake stars are chosen, and (ii) the fewer real stars a participant assigns to her partner, the more fake stars are selected.

Result 1 *Participants use fake stars, and the amount of fake stars and the fraction of cheaters decreases with increasing costs. The usage of fake stars increases over time, but only when cheating is free. The fewer real stars a participant gets, the more she makes use of fake stars.*

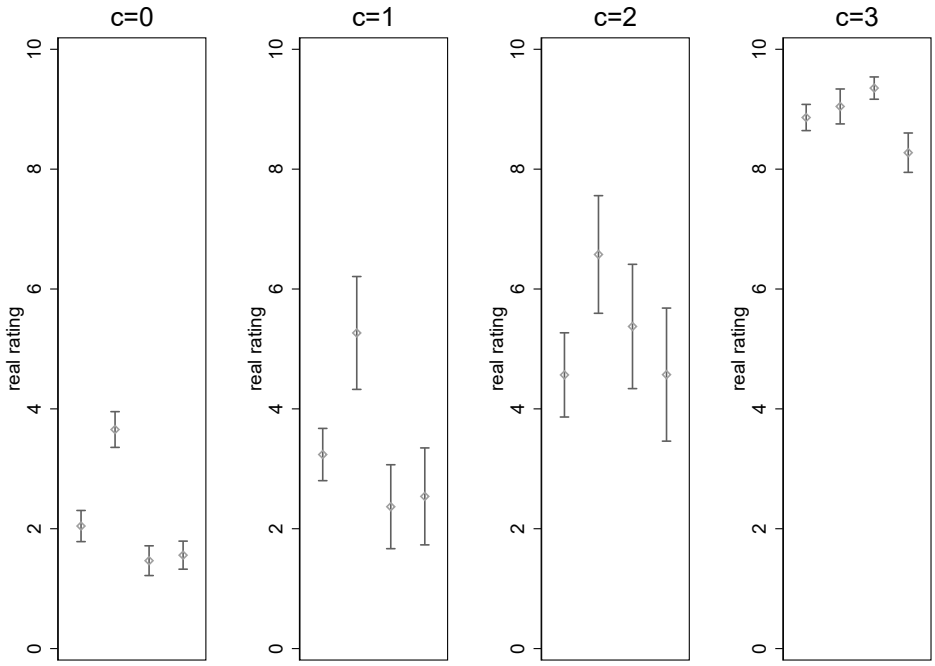


Figure 2: The graphs show the mean values and 95% confidence intervals of the real rating assigned to the partner (y-axis). Each panel shows means and confidence intervals for a given contribution ($c = 0$, $c = 1$, $c = 2$, $c = 3$) dependent on the treatment (x-axis). Order of treatments (x-axis): *Control*, *No costs*, *Low costs* and *High costs*.

Next, we analyze evaluation behavior, which refers to the rating a participant receives from her interaction partner (i.e., fake stars are not included). We analyze if real ratings correlate with behavior in the four treatments. Pooled over all participants and periods, the rank correlation between the partner's contribution and the real rating the partner received within the same period is 0.715 in *Control*, 0.5192 in *No costs*, 0.822 in *Low costs*, and 0.688 in *High costs*. Spearman's correlation tests reveal that all correlations are significant and positive ($p < 0.001$). Hence, real ratings are noisy but are good predictors of behavior. All correlations are significantly different ($p < 0.001$) except the correlations in *Control* and *High costs* ($p = 0.316$).[11] The correlation of contributions and ratings differ significantly between treatments: With increasing costs for fake stars, correlations are higher.

[11]The p-values from the comparisons of correlations between treatments are bootstrapped with 1000 repetitions.

	(5)	(6)	(7)	(8)	(9)
c_j	2.195***	2.256***	2.137***	2.149***	2.160***
	(0.072)	(0.132)	(0.076)	(0.075)	(0.074)
No costs	1.335***	1.833***	1.039***	1.020**	0.338
	(0.292)	(0.440)	(0.311)	(0.310)	(0.398)
Low costs	0.084	-0.398	-0.126	-0.151	-0.161
	(0.203)	(0.341)	(0.216)	(0.214)	(0.298)
High costs	-0.415	-0.331	-0.454	-0.450	-0.630
	(0.255)	(0.346)	(0.266)	(0.265)	(0.374)
c_i	0.015	-0.009	-0.029	-0.008	0.004
	(0.056)	(0.057)	(0.059)	(0.059)	(0.059)
r_i			0.020	0.029	0.019
			(0.026)	(0.026)	(0.026)
r_j			0.070***	0.071***	0.064**
			(0.021)	(0.021)	(0.020)
Period				0.029*	0.001
				(0.015)	(0.023)
No costs×c_j		-0.506*			
		(0.204)			
Low costs×c_j		0.314			
		(0.174)			
High costs×c_j		-0.061			
		(0.189)			
No costs×Period					0.089*
					(0.043)
Low costs×Period					0.006
					(0.033)
High costs×Period					0.021
					(0.039)
Constant	1.882***	1.830***	1.543***	1.201***	1.496***
	(0.190)	(0.259)	(0.248)	(0.294)	(0.320)
N	4590	4590	4284	4284	4284
$R^2(within)$	0.533	0.542	0.532	0.533	0.534
$R^2(between)$	0.425	0.430	0.429	0.430	0.428
$R^2(overall)$	0.505	0.513	0.503	0.504	0.505

Table 3: Real ratings as a function of the partner's contribution c_j, the participant's own contribution c_i, treatment dummies *No costs*, *Low costs* and *High costs*. r_i is the participant's own and r_j is the partner's displayed rating. *Control* is the baseline treatment. Regressions are random effects regressions with robust standard errors. In regression (1), N is 4590 and therewith 306 cases higher than in the remaining regressions because observations from the very first period are also considered. In regressions (6) - (8), the consideration of r_i and r_j excludes observations from the first period because no previous rating can be considered (therefore, r_i and r_j are not available in the first period). Significance levels: $*$ for $p < 0.05$, $**$ for $p < 0.01$, $***$ for $p < 0.001$.

In all treatments, low contributions tend to receive bad real ratings and high contributions receive good real ratings (compare Figure 2). However, in *No costs*, real ratings for low contributions are higher and therefore less distinctive. Participants receive on average almost 4 (5) stars for contributions of $c = 0$ ($c = 1$), which is significantly more than what counterparts from the other treatments receive (compare the two sub-panels in Figure 2 furthest to the left).

The analysis of correlations and the visual inspection are confirmed by random effects regressions, displayed in Table 3, in which we regress the real rating assigned to the partner on the partner's contribution (c_j), controlling for the participant's own (r_i) and the partner's displayed rating (r_j), the participant's own contribution (c_i), time effects and interactions. We include indicator variables for treatments, taking *Control* as the baseline category.

Regression (5) - (9) in Table 3 corroborate that with increasing contributions (c_j), participants assign more real stars. The coefficient of c_j is positive and significant. The regressions also reveal that only the dummy variable for *No costs* is significantly different to the benchmark case *Control*. Participants assign significantly more real stars when cheating is free. Regression (6) shows that the interaction variable *No costs*$\times c_j$ is significantly negative. The relationship between contribution and real stars is lower when there are no costs for fake stars. Regressions (8) and (9) corroborate that the positive time effect is only driven by evaluation behavior in treatment *No costs* (the coefficient of the interaction *No costs*$\times$Period is positive and significant).

Result 2 *High contributions lead to better real ratings than low contributions in all treatments. However, this positive relationship is weaker when fake stars are free, where real ratings inflate over time.*

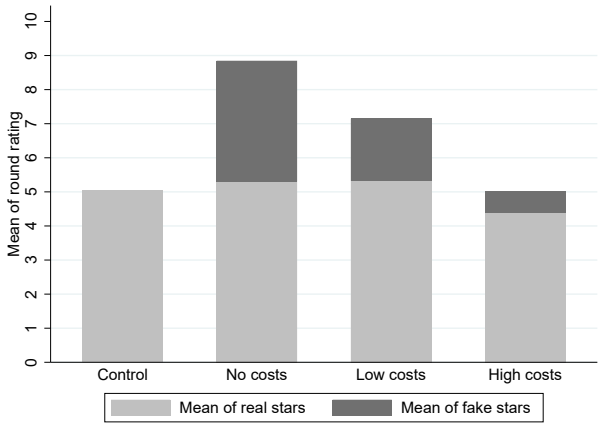


Figure 3: The bar chart shows means of round ratings, means of real stars and means of fake stars in the four treatments *Control*, *No costs*, *Low costs* and *High costs*.

In a final step, we depict the round rating, which is composed of the fake stars and real stars from one period. Figure 3 summarizes the means of this variable. It can be taken from Figure 3 that the mean of real stars does not differ between treatments, whereas the mean of fake stars differs between treatments (compare the analysis above). Moreover, in *No costs*, the mean of the round rating is about 9 stars, which shows that round ratings in this treatment are almost always near the maximum of 10 stars.

In order to analyze whether the round rating is informative about past behavior, we look at the correlation between a participant's own contribution and round rating. Pooled over all participants and periods, the rank correlation between a participant's own contribution and round rating within the same period is 0.715 in *Control*, 0.200 in *No costs*, 0.583 in *Low costs*, and 0.660 in *High costs*. Spearman's correlation tests reveal that all correlations are significant and positive ($p < 0.001$). A series of difference-tests with bootstrapped standard errors reveal that all correlations are significantly different with p-values of $p = 0.038$ when comparing *Control* and *High costs*, $p = 0.012$ when comparing *Low costs* and *High costs* and with $p < 0.001$ for all remaining treatment comparisons.

The analysis of the correlation between round ratings and contributions reveals that the reliability of ratings is highest in *Control* and lowest in *No costs*. Moreover, when the costs of cheating increase, the reliability (correlation) increases significantly. Another way to look at this correlation is presented in Figure 4, which is the equivalent of Figure 2, but with round ratings (i.e., fake stars plus real stars) instead of only the real ratings.

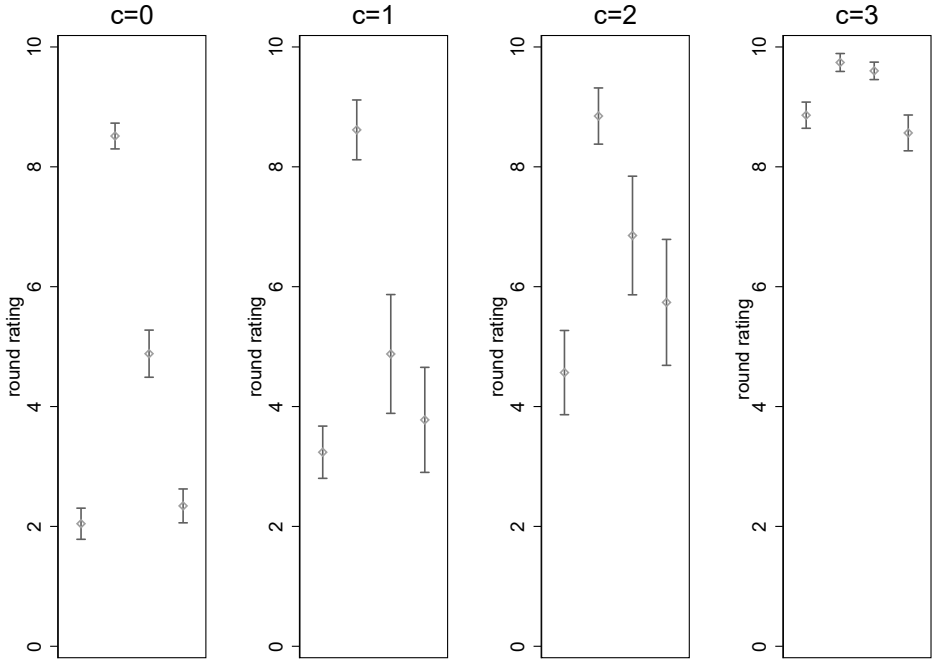


Figure 4: The graphs show the mean values and 95% confidence intervals of the round rating at the end of a period (y-axis). Each panel shows means and confidence intervals for a given contribution ($c = 0$, $c = 1$, $c = 2$, $c = 3$) dependent on the treatment (x-axis). Order of treatments (x-axis): *Control*, *No costs*, *Low costs* and *High costs*.

Remember that in the experiment, participants are only informed about the displayed rating, which refers to the average round rating of the last three periods. In this subsection 3.1, our analysis concerning the reliability of ratings has instead focused on the round rating. Such an analysis is more accurate than analyzing the displayed rating because for each single contribution, the resulting real rating, fake stars and round rating can be studied. Moreover, if round ratings are informative in each period, then the average round rating (i.e., the displayed rating) also has to be informative about contribution behavior. Hence, we conclude the following:

Result 3 *Displayed ratings are most reliable in Control and least reliable in No Costs. With increasing costs for fake stars, the reliability of displayed ratings increases as well. Ratings in High costs are nearly as informative about behavior as in Control.*

In the following subsections, we focus our analysis on the displayed rating (r_i and r_j) in order to analyze contributions, efficiency and expectations.

3.2 Contributions and Efficiency

In this section, we analyze contributions and efficiency. Table 4 summarizes some descriptive statistics and the distribution of contributions by treatment. We observe that in the treatment *No costs*, the mean of contributions (0.95) is much lower than in *Control* (1.43), *Low costs* (1.52) and *High costs* (1.32), where contributions are on a similar level. The distribution of contributions between treatments (compare Table 4) indicates that the differences in means can be explained by substantially more free-riders ($c = 0$) and less participants with full contributions ($c = 3$) in *No costs* than in the other treatments.

	Mean	Distribution of Contributions				
Treatment	(sd)	0	1	2	3	Total
Control	1.43	579	147	90	534	1350
	(1.38)	42.89%	10.89%	6.67%	39.56%	100.00%
No costs	0.95	679	60	59	282	1080
	(1.31)	62.87%	5.56%	5.46%	26.11%	100.00%
Low costs	1.52	484	49	48	499	1080
	(1.44)	44.81%	4.54%	4.44%	46.20%	100.00%
High costs	1.32	549	63	42	426	1080
	(1.42)	50.83%	5.83%	3.89%	39.44%	100.00%

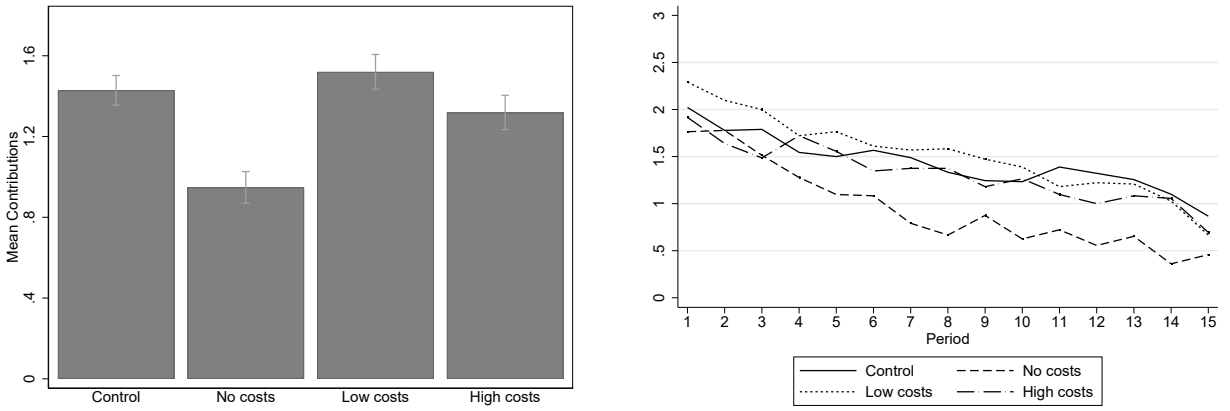
Table 4: Mean and distribution of contributions by treatment (data pooled across periods).

Figure 5b shows that in all treatments, contributions decrease over time. However, in treatment *No costs* contributions are always lower than in the remaining treatments, except for the second period. The latter finding is important because in addition to our random allocation into treatments, participants in *No costs* seem to not differ in their general propensity to contribute. Interestingly, contributions decrease smoothly over time and the end-game effect is rather weak.[12]

Pooled over periods, the difference in contributions between *Control* and *No costs* is 0.48 (1.43-0.95) (see Table 4 and Figure 5a). A Mann-Whitney-test with session averages shows that this difference is significant ($p = 0.028$, $z = -2.205$). A series of Mann-Whitney-tests also reveal that contributions in *Control* do not differ from contributions both in *Low costs* and *High costs*.[13]

[12]This finding is in line with several public good games with social approval. Without social approval, most repeated public good experiments find a sharp drop in contributions in the last periods. Compare Greiff & Paetzel (2020), Greiff & Paetzel (2015) and Pan & Houser (2017) for further details.

[13]When we compare contributions in *Control* and contributions in *Low costs*, we find a difference of 0.09 (1.52-1.43), which is insignificant ($p = 1$, $z = 0$); when we compare contributions in *Control* and contributions in *High costs*, we find a difference of 0.11 (1.43-1.32), which is also insignificant ($p = 0.9021$, $z = -0.123$).



(a) Average contributions by treatment pooled over periods.

(b) Contributions over time; data points are means of absolute contributions for each period.

Figure 5: Subfigure (a) shows average contributions. Subfigure (b) presents contributions over time.

Contributions are about 50% lower when cheating is possible and free. The remaining treatment comparisons are delegated to the following regression analysis.

So far, we have investigated ratings and contributions separately. In the following, we will analyze how displayed ratings (r_i and r_j) affect contributions. Remember that displayed ratings are averages of round ratings, including fake stars and real stars.

Regressions in Table 5 show how contributions are affected by the information contained in the participant's own displayed rating r_i and the partner's displayed rating r_j. While regressions (10) to (13) are treatment-specific, regressions (14) to (16) are based on the pooled data from all four treatments. In all regressions, both the participant's own and the partner's displayed ratings have a significant and positive effect on contributions, which makes the information provided by the reputation system reliable even when cheating is free. It can also be taken from all regressions that there is a negative time trend when we control for displayed ratings (except for *Control*).

In regressions (14) to (16), we include indicator variables for treatments, taking *Control* as the baseline condition. In regression (14), only the indicator variable for *No costs* is significantly negative. Hence, contributions are significantly lower in *No costs* than in the other treatments. A Wald-test reveals that in regression (15) there is no significant difference between the indicator variables for *Low costs* and *High costs* ($p = 0.2207$). Regression (15) shows that when displayed ratings are considered, the indicator variable for *Low costs* is also significantly negative. Only *High costs* shows contributions similar to contributions in *Control*.

Note that in regressions (10), (12) and (13), the coefficient of the partner's displayed rating (r_j) is about twice as large as in regressions (11). This suggests that when cheating is free, participants react less strongly to the information about the partner's displayed rating. Regression (16) shows that this effect is statistically significant. Regression (16) includes interaction terms between the indicator variables for the treatments and r_j. The coefficient on *No costs*$\times\ r_j$ is significantly negative. A Wald-test reveals that there is no significant difference in the coefficients of *Low costs*$\times\ r_j$ and *High costs*$\times\ r_j$. Hence, the marginal effect of r_j is significantly lower in *No costs* compared to the other treatments. This reveals that with free manipulation, participants rely less on the information they receive.

	(10)	(11)	(12)	(13)	(14)	(15)	(16)
Treatment(s)	*Control*	*No costs*	*Low costs*	*High costs*	(All)	(All)	(All)
r_i	0.138***	0.038*	0.069***	0.084***		0.063***	0.063***
	(0.014)	(0.018)	(0.017)	(0.018)		(0.009)	(0.009)
r_j	0.138***	0.073***	0.127***	0.132***		0.125***	0.137***
	(0.014)	(0.017)	(0.019)	(0.018)		(0.009)	(0.014)
No costs					-0.481***	-1.168***	-0.655***
					(0.137)	(0.137)	(0.166)
Low costs					0.091	-0.348**	-0.336*
					(0.151)	(0.131)	(0.163)
High costs					-0.109	-0.111	-0.043
					(0.145)	(0.128)	(0.135)
No costs$\times r_j$							-0.063**
							(0.022)
Low costs$\times r_j$							-0.005
							(0.023)
High costs$\times r_j$							-0.013
							(0.022)
Period	-0.014	-0.091***	-0.065***	-0.020*	-0.078***	-0.051***	-0.050***
	(0.008)	(0.012)	(0.011)	(0.009)	(0.005)	(0.005)	(0.005)
Constant	0.055	0.683**	0.540**	0.302*	2.052***	0.829***	0.763***
	(0.122)	(0.210)	(0.206)	(0.154)	(0.103)	(0.106)	(0.105)
N	1260	1008	1008	1008	4590	4284	4284
$R^2(within)$	0.176	0.172	0.236	0.209	0.106	0.195	0.198
$R^2(between)$	0.708	0.091	0.439	0.469	0.051	0.333	0.327
$R^2(overall)$	0.314	0.113	0.293	0.275	0.081	0.237	0.237

Table 5: Contributions as a function of displayed ratings in periods 2 to 15. Regressions are random effects regressions with robust standard errors in parentheses. r_i is the participant's own rating and r_j is the partner's displayed rating. In regressions (14)-(16) *No costs*, *Low costs* and *High costs* are treatment dummies and *Control* is the baseline treatment. Regression (10)-(13) are regressions for each single treatment. Significance levels: $*$ for $p < 0.05$, $**$ for $p < 0.01$, $***$ for $p < 0.001$.

Result 4 *Contributions are significantly lower in No costs than in the control treatment. When cheating comes at a cost, however, contributions remain on the same level as in Control. Participants also react more strongly to the partner's displayed rating when there are costs of manipulation.*

As a final step in this subsection, we will analyze efficiency. If we define efficiency as aggregated outcomes and ignore differences in costs between reputation systems, efficiency should correspond to contributions. However, the clean measurement of efficiency also considers costs. Table 6 summarizes averages of gross profits, average costs and resulting net profits for each treatment. A series of Mann-Whitney-tests reveals the following: (i) Net profits in *Control* are significantly higher than in *No costs* and *High costs* ($p < 0.001$), but are not different to net profits in *Low costs* ($p = 0.1198$); (ii) net profits are higher in *Low costs* than in *High costs* ($p = 0.0731$), (iii) net

Treatments	*Control*	*No Costs*	*Low costs*	*High costs*
mean gross profit (sd)	12.86 (3.89)	11.90 (4.04)	13.04 (3.97)	12.64 (3.91)
mean fake costs (sd)	0	0	0.36 (0.70)	0.37 (1.21)
mean net profit (sd)	12.86 (3.89)	11.90 (4.04)	12.68 (3.93)	12.27 (4.03)
N	1350	1080	1080	1080

Table 6: Mean (sd) of gross profit, costs and net profit by treatment.

profits are higher in *High costs* than in *No costs* ($p = 0.0195$) and (iv) net profits are significantly higher in *Low costs* than in *No costs* ($p = 0.0010$). The analysis of net profits shows that there is inefficiency both in statistically and economically relevant terms when manipulation opportunities are free. The inefficiency caused by free fake stars is about 7.5% ((12.86-11.90)/12.86). We observe that this inefficiency can be reduced with costs for fake stars.[14]

Result 5 *Manipulation opportunities cause inefficiency when fake stars are free. This inefficiency is reduced by introducing costs of manipulation. In Low costs, efficiency is as high as in Control.*

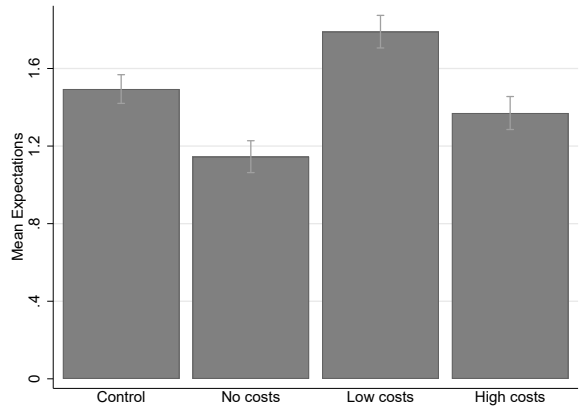
3.3 Expectations

In this Subsection 3.3 we analyze expectations. The analysis of expectations dissects how participants decide their contributions based on both displayed ratings and opportunities to cheat. In line with Greiff & Paetzel (2016), we argue that cooperation is driven by a preference for conditional cooperation. Based on the information provided by the reputation system, a conditional cooperator forms expectations about her partner's contribution, and, based on this expectation, the participant chooses her contribution. A conditional cooperator who expects her partner to choose a high (low) contribution will also choose a high (low) contribution. We analyze both how expectations rely on displayed ratings and how differences in contributions can be explained.

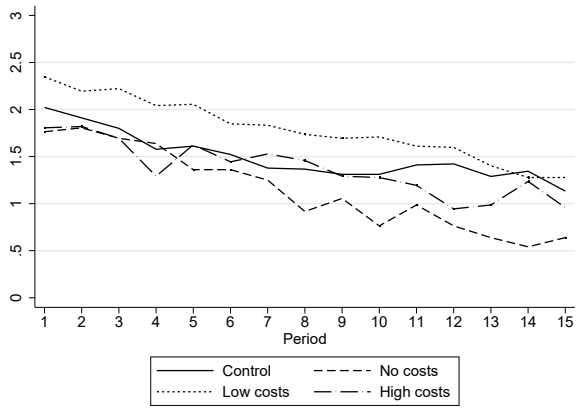
As can be taken from Figure 6, expectations are very similar to contributions pooled over periods and over time. In particular, in *No costs* participants expect their partner to contribute significantly less than in the remaining treatments. Expectations about contributions are always slightly above actual contributions. Conditional cooperation necessarily results in a positive correlation between expectations and contributions. In all treatments, the spearman rank correlation is significant and on a high level (0.6289 in *Control*, 0.6086 in *No costs*, 0.6752 in *Low costs* and 0.6724 in *High costs*). This finding is perfectly in line with previous work on conditional cooperation (for example Fischbacher et al. 2001, Fischbacher & Gächter 2010, Dorner et al. 2020). The correlations are not significantly different from each other, except for the correlations of *Low costs* and *No costs*, which are both significantly higher than the correlation in *No costs* ($p = 0.040$ and $p = 0.056$).

Further regressions with expectations as the dependent variable reveal that participants rely less on the partner's evaluation in *No costs*, which implies that conditional cooperation is lower.

[14] One might argue that the measurement of efficiency in our experiment depends on our specific parametrization. We acknowledge that the quantitative differences might be different with another parametrization. However, the qualitative treatment differences should remain stable even with a different parametrization of payoffs and different costs for fake stars.



(a) Average expectations about contributions of the new partner by treatment (pooled over periods).



(b) Expectations over time; data points are means of absolute contributions for each period.

Figure 6: Subfigure (a) shows average expectations. Subfigure (b) presents expectations over time.

These results are in line with the results of the regressions for contributions, displayed in Table 5). Instead of providing these additional regressions, Figure 7 shows how a participant forms their expectations based on both their own displayed rating (y-axis) and the partner's displayed rating (x-axis).

The different shades of gray in Figure 7 represent expectations for all combinations of the participant's own and the partner's displayed ratings, with darker colors corresponding to higher expectations. Squares in light gray correspond to combinations of the participant's own and the partner's displayed ratings, for which the average expectation (over all participants with this particular combination of displayed ratings) lies in the interval $[0, 1]$. Squares in medium gray correspond to expectations in the interval $(1, 2]$. Black squares correspond to the highest possible expectation: an expectation in the interval $(2, 3]$.

When comparing the black areas between treatments in Figure 7, it becomes clear that in *Control*, the black area in the upper-right corner is "largest". Expectations in *Control* are high when both the participant's own and the partner's displayed ratings are above 6 stars. Expectations in *Low costs* are generally more pessimistic. Only when the participant's own and the partner's displayed ratings are very high (8-10 stars on average) do the expectations lie in the interval $(2, 3]$. Figure 7 shows that in treatments with the option to cheat (*No costs*, *Low costs* and *High costs*), the black area increases with increasing manipulation costs. We interpret a higher expectation as being more "optimistic" due to the partner's contribution.

Interestingly, in *No costs* expectations are high when the partner's displayed rating is between 2-4 stars. Participants in *No costs* seem to anticipate high numbers of fake stars and do not trust high displayed ratings. Instead, they expect that a low but positive displayed rating can be interpreted as a positive signal that can be trusted: In *No costs*, the "feedback language" of the reputation system seems to change compared to the other treatments.

Result 6 *With increasing costs for fake stars, participants rely more on the displayed ratings and, subsequently, expectations increase.*

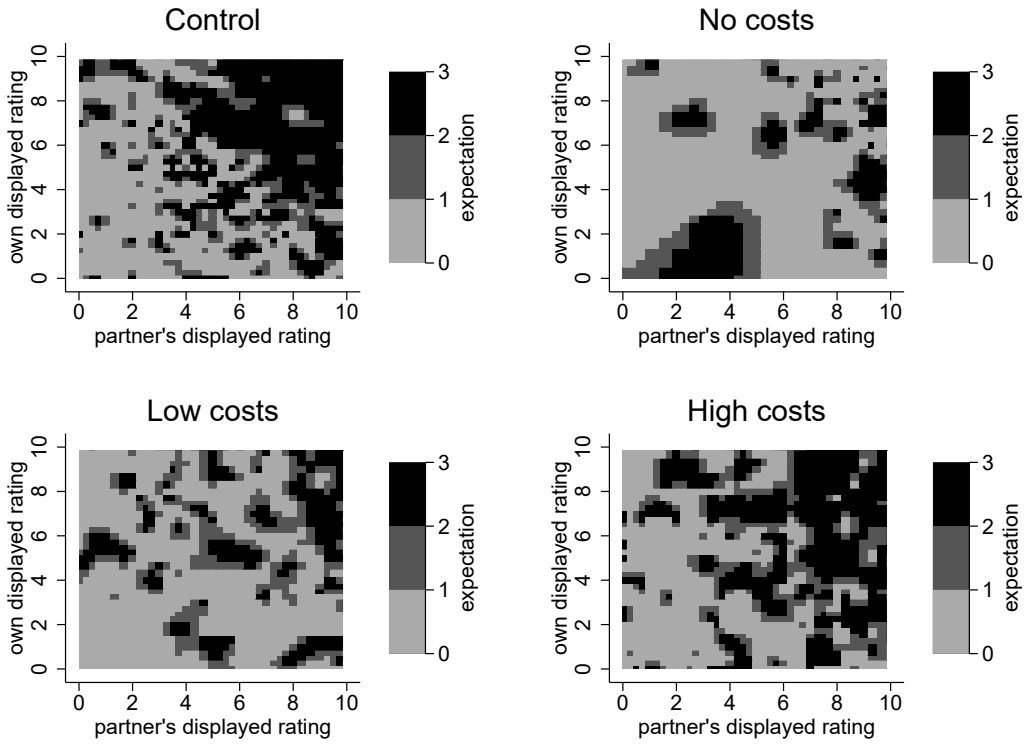


Figure 7: Expectations as a function of the participant's own displayed rating and the partner's displayed rating. Figures are created using the Stata function *twoway contour* with the *heatmap* option. Missing entries for expectations were interpolated based on the values from neighboring cells.

4 Conclusion

Recent media reports and sector inquiries highlight the importance of developing a better understanding of how fake reviews impact the functioning of reputation systems and whether they lead to market inefficiency (e.g The Guardian 2019, German Federal Cartel Office 2020). In this work, we studied the impact of manipulated feedback on reputation systems. Using a repeated public goods experiment, we investigated how participants make use of manipulation opportunities, how manipulated feedback influences the reliability of average ratings and whether introducing costs of manipulation has a positive effect on the efficiency of the reputation system. We conducted three treatments in which the cost of manipulation varied: *No costs*, *Low costs* and *High costs*. In addition, we conducted a control treatment where cheating was not possible.

In our *Main Hypothesis*, we proposed that when cheating is possible and there are no costs of manipulation, ratings will not be informative about the contribution behavior of participants. We expected contributions and overall efficiency to be lower in *No costs* than in the control treatment. Our evidence confirms our *Main Hypothesis*: The reliability of ratings decreases significantly in *No costs* compared to *Control*. Expectations, contributions to the public good and overall efficiency are also significantly lower in *No costs* than in *Control*.

Furthermore, we conjectured that introducing costs of manipulation in the treatments *Low costs* and *High costs* would have a positive influence on the reliability of displayed ratings compared to

the *No costs* treatment. We also expected contributions and overall efficiency to be higher in *Low costs* and *High costs* than in *No costs*. We found causal evidence supporting this conjecture: The reliability of displayed ratings increases with increasing manipulation costs, and contributions as well as efficiency are generally higher when manipulation costs exist. While the reliability of ratings is higher in *High costs* than in *Low costs*, contributions are on a similar level in both treatments. The net efficiency in *Low costs* is not significantly different from the net efficiency in the control treatment.

Our work contributes to recent research tackling the analysis of fake reviews in reputation systems (e.g. Ott et al. 2012, Mayzlin et al. 2014, Luca & Zervas 2016). In our controlled laboratory experiment, the potential problems of reputation systems such as selection bias and reciprocal evaluation behavior are excluded. In addition, the experimental environment enables us to detect fake ratings unambiguously. In contrast to field studies, utilizing an experiment also allows us to analyze the complete transmission channel of manipulated ratings from evaluation behavior to cheating behavior, expectation formation and contribution behavior. Hence, we believe that our experimental approach offers insights into the effects of fake reviews on reputation systems and efficiency that can complement field studies on the subject.

Our study also contributes to the literature on cheating behavior (e.g. Fischbacher & Föllmi-Heusi 2013, Gneezy et al. 2018, Abeler et al. 2019). To the best of our knowledge, this study is the first experiment that enables participants to manipulate feedback by others and the first to analyze cheating behavior in a repeated public good game with reputation systems. Although our experiment differs fundamentally from previous static cheating experiments without strategic interactions, we also find that subjects cheat, but not to the full extent. However, as opposed to previous literature, we find that the net outcome of a lie significantly matters for the decision to cheat. Additionally, we observe that in the treatment with no cheating costs, the preference for "having a good reputation" seems to be more important than a preference for being honest.

Our findings suggest that the efficiency of a reputation system will be low if online platforms do nothing in order to stop fake reviews. When there are costs of manipulation, for example because online platforms use algorithms or other devices to remove fake content, the reliability of ratings is higher and the efficiency of the rating system increases. Our findings are interesting for managers of online platforms, since our results suggest that firms do not have to detect all fake reviews. Rather, our experimental evidence implies that low manipulation costs are sufficient for a reputation system to work effectively.

References

Abeler, J., Nosenzo, D. & Raymond, C. (2019), 'Preferences for truth-telling', *Econometrica* **87**(4), 1115–1153.

Athey, S. & Luca, M. (2019), 'Economists (and economics) in tech companies', *Journal of Economic Perspectives* **33**(1), 209–30.

Ba, S. & Pavlou, P. A. (2002), 'Evidence of the effect of trust building technology in electronic markets: Price premiums and buyer behavior', *MIS quarterly* **26**(3), 243–268.

Bock, O., Baetge, I. & Nicklisch, A. (2014), 'hroot: Hamburg registration and organization online tool', *European Economic Review* **71**, 117–120.

Bolton, G. E., Kusterer, D. J. & Mans, J. (2019), 'Inflated Reputations: Uncertainty, Leniency, and Moral Wiggle Room in Trader Feedback Systems', *Management Science* **65**(11), 5371–5391.

Bolton, G., Greiner, B. & Ockenfels, A. (2013), 'Engineering Trust: Reciprocity in the Production of Reputation Information', *Management Science* **59**(2), 265–285.

CBC News (2015), 'Bell hit with $1.25m fine for planting 4-star reviews for phone apps', https://www.cbc.ca/news/business/bell-hit-with-1-25m-fine-for-planting-4-star-reviews-for-phone-apps-1.3271222. Accessed: 2020-10-28.

Charness, G., Masclet, D. & Villeval, M. C. (2014), 'The dark side of competition for status', *Management Science* **60**(1), 38–55.

Chen, Y., Cramton, P., List, J. A. & Ockenfels, A. (2020), 'Market design, human behavior, and management', *Management Science* . Published online in Articles in Advance 09 Sep 2020. https://doi.org/10.1287/mnsc.2020.3659.

Dellarocas, C. (2003), 'The Digitization of Word of Mouth: Promise and Challenges of Online Feedback Mechanisms', *Management Science* **49**(10), 1407–1424.

Dorner, V., Giamattei, M. & Greiff, M. (2020), 'The Market for Reviews: Strategic Behavior of Online Product Reviewers with Monetary Incentives', *Schmalenbach Business Review* **72**, 397–435.

Federal Trade Commission (2019), 'Devumi, owner and CEO settle FTC charges they sold fake indicators of social media influence; Cosmetics firm Sunday Riley, CEO settle FTC charges that employees posted fake online reviews at CEOs direction', https://www.ftc.gov/news-events/press-releases/2019/10/devumi-owner-ceo-settle-ftc-charges-they-sold-fake-indicators. Accessed: 2020-10-28.

Fischbacher, U. (2007), 'z-tree: Zurich toolbox for ready-made economic experiments', *Experimental economics* **10**(2), 171–178.

Fischbacher, U. & Föllmi-Heusi, F. (2013), 'Lies in disguise - an experimental study on cheating', *Journal of the European Economic Association* **11**(June), 525–547.

Fischbacher, U. & Gächter, S. (2010), 'Social preferences, beliefs, and the dynamics of free riding in public goods experiments', *American economic review* **100**(1), 541–56.

Fischbacher, U., Gächter, S. & Fehr, E. (2001), 'Are People Conditionally Cooperative? Evidence from a Public Goods Experiment', *Economic Letters* **71**(3), 397–404.

Fradkin, A., Grewal, E. & Holtz, D. (2018), 'The determinants of online review informativeness: Evidence from field experiments on airbnb', *Available at SSRN 2939064* .

German Federal Cartel Office (2020), 'Fake and manipulated user reviews for online purchases - Bundeskartellamt provides background information and solution approaches', https://www.bundeskartellamt.de/SharedDocs/Meldung/EN/Pressemitteilungen/2020/06_10_2020_SU\%20Nutzerbewertungen.html. Accessed: 2020-10-28.

Gneezy, U., Kajackaite, A. & Sobel, J. (2018), 'Lying aversion and the size of the lie', *American Economic Review* **108**(2), 419–53.

Greiff, M. & Paetzel, F. (2015), 'Incomplete information strengthens the effectiveness of social approval', *Economic Inquiry* **53**(1), 557–573.

Greiff, M. & Paetzel, F. (2016), 'Second-order beliefs in reputation systems with endogenous evaluations–an experimental study', *Games and Economic Behavior* **97**, 32–43.

Greiff, M. & Paetzel, F. (2020), 'Information about average evaluations spurs cooperation: An experiment on noisy reputation systems', *Journal of Economic Behavior and Organization* **180**, 334–356.

Hu, N., Pavlou, P. A. & Zhang, J. J. (2017), 'On self-selection biases in online product reviews', *MIS Quarterly* **41**(2), 449–471.

Hui, X., Saeedi, M., Shen, Z. & Sundaresan, N. (2016), 'Reputation and Regulations: Evidence from eBay', *Management Science* **62**(12), 3604–3616.

Karaman, H. (2020), 'Online review solicitations reduce extremity bias in online review distributions and increase their representativeness', *Management Science* . Published online in Articles in Advance 29 Oct 2020. https://doi.org/10.1287/mnsc.2020.3758.

Luca, M. (2017), 'Designing online marketplaces: Trust and reputation mechanisms', *Innovation Policy and the Economy* **17**(1), 77–93.

Luca, M. & Zervas, G. (2016), 'Fake It Till You Make It: Reputation, Competition, and Yelp Review Fraud', *Management Science* **62**(12), 3412–3427.

Masclet, D. & Pénard, T. (2012), 'Do reputation feedback systems really improve trust among anonymous traders? An experimental study', *Applied Economics* **44**(35), 4553–4573.

Mayzlin, D., Dover, Y. & Chevalier, J. (2014), 'Promotional reviews: An empirical investigation of online review manipulation', *American Economic Review* **104**(8), 2421–55.

Necker, S. & Paetzel, F. (2020), 'Bad losers? bad winners? the effect of the quality of competitors on dishonesty and effort in repeated competitions', *SSRN Working Paper No. 3506551* .

New York State Office of the Attorney General (2013), 'A.G. Schneiderman announces agreement with 19 companies to stop writing fake online reviews and pay more than $350,000 in fines', https://ag.ny.gov/press-release/2013/ag-schneiderman-announces-agreement-19-companies-stop-writing-fake-online-reviews. Accessed: 2020-10-28.

Ott, M., Cardie, C. & Hancock, J. (2012), Estimating the prevalence of deception in online review communities, *in* 'Proceedings of the 21st international conference on World Wide Web', pp. 201–210.

Pan, X. & Houser, D. (2017), 'Social approval, competition and cooperation', *Experimental Economics* **20**(2), 309–332.

Schoenmueller, V., Netzer, O. & Stahl, F. (2020), 'The polarity of online reviews: Prevalence, drivers and implications', *Journal of Marketing Research* **57**(5), 853–877.

Stahl, D. O. (2013), 'An experimental test of the efficacy of a simple reputation mechanism to solve social dilemmas', *Journal of Economic Behavior & Organization* **94**, 116–124.

Stiftung Warentest (2020), 'Fake-Bewertungen: Wie Verkäufer mit gekauftem Lob Kunden manipulieren', https://www.test.de/Fake-Bewertungen-Wie-Verkaeufer-mit-gekauftem-Lob-Kunden-manipulieren-5401497-0/. Accessed: 2020-10-28.

Tadelis, S. (2016), 'Reputation and Feedback Systems in Online Platform Markets', *Annual Review of Economics* **8**(1), 321–340.

The Guardian (2019), 'TripAdvisor is failing to stop fake hotel reviews, says Which?', https://www.theguardian.com/travel/2019/sep/06/tripadvisor-failing-to-stop-fake-hotel-reviews-which. Accessed: 2020-10-28.

A Experimental Instructions

Welcome to the experiment and thank you for participating!

Please read the instructions carefully. Do not talk to your neighbors at any point during the experiment. If you have any questions, please raise your hand. One of the experimenters will come to you and answer your questions privately. Following this rule is very important. Otherwise the results of this experiment will be significantly compromised from a scientific point of view.[15]

Please take your time reading the instructions and making your decisions. You are not able to influence the duration of the experiment by rushing through your decisions, because you will always have to wait until all other participants have made their decisions. The experiment is completely anonymous. At no time during or after the experiment will the other participants know which role you were assigned to and how much you have earned.

You will receive a show-up fee of 5 euros for your participation. Depending on your decisions and the decisions of the other participants you can earn additional money up to 23 euros. You will be paid individually, privately and in cash after the experiment. The expected duration of the experiment is 90 minutes. The exact procedure of the experiment will be described in the following.

The experiment consists of 15 rounds which all follow the same procedure. In each round participants will be randomly and repeatedly assigned to groups of two members. Your payoff will be determined solely by your own decisions and the decisions of your partner. The decisions of the other groups do not affect your payment. You will not encounter the same partner in subsequent rounds.

Within a round

You and another participant will form a group of two in each round. Both members will be asked about their expectations regarding the decision of the other member, will make a decision on their own and will evaluate the decision of the other member. Afterwards, there is the opportunity to improve the evaluation that you received from your partner in return for money. This completes a round. The resulting decision combination from your and the other member's decision determines your payoff.

		the other participant chooses			
		A	**B**	**C**	**D**
you choose	**A**	10 | 10	13 | 9	16 | 8	19 | 7
	B	9 | 13	12 | 12	15 | 11	18 | 10
	C	8 | 16	11 | 15	14 | 14	17 | 13
	D	7 | 19	10 | 18	13 | 17	16 | 16

Figure 8: Payoff-bimatrix.

The associated payoffs (in euros) are listed in Figure 8. The possible costs of improving your own evaluation are not included in Figure 8. Figure 8 is also shown on the decision screen and contains every possible decision you can make in its row head. The possible decisions of your partner are listed in the column head. The corresponding payoffs for you and your partner can be found in the cells in which the rows and columns intersect. The amount on the left of the vertical bar is your payment, the amount on the right of the vertical bar is the payment of your partner.

Starting with the second round, you will be informed at the beginning of each round about how your new partner has been evaluated in previous rounds. However, you will not be informed about whether the original evaluation of your new partner has been changed or not. Your partner will be informed about how you have been evaluated. Your partner will also not be informed about whether you changed your original evaluation or not.

- At the beginning of the second round, you will be informed about how your partner has been evaluated in the first round. Your partner will be informed about how you have been evaluated in the first round.

[15] The original instructions were in German. This is an example for the treatment *Low costs*. The instructions for the other treatments are available upon request.

- At the beginning of the third round, you will be informed about how your partner has been evaluated in the previous two rounds. The average rating of the first two rounds of your partner will be displayed to you. Your partner will be informed about how you have been evaluated in the previous two rounds. Your average rating of the first two rounds will be displayed to her.
- At the beginning of the fourth round, you will be informed about how your partner has been evaluated in the previous three rounds. The average rating of the first three rounds of your partner will be displayed to you. Your partner will be informed about how you have been evaluated in the previous three rounds. Your average rating of the first three rounds will be displayed to her.

Before you decide, you will be asked what decision you expect your partner to make. Afterwards you and your partner will decide at the same time. After that, both of you will get informed about your payoffs.

After you have been informed about your payoff, you will be able to evaluate your group members' decision. For this purpose, your own decision and payoff as well as the decision and payoff of your partner will be displayed. To evaluate, you can grant up to 10 stars, where 0 stars is the worst possible and 10 stars is the best possible evaluation. In the next step, your partner will be informed about how you evaluated her and you will be informed about how you have been evaluated by your partner.

On the next decision screen, you will be able to change the evaluation that you received from your partner by buying additional stars. Each additional star costs 20 cents. In each period, you can buy as many additional stars until you have reached the best possible evaluation of 10 stars.

Calculation of your final payoff

Your final payoff consists of three parts:

(i) The show-up fee of 5 euros.

(ii) The second part of your payoff is determined by the payoff resulting from your and your partner's decision in a round (displayed in Figure 8) minus the costs for stars that you bought in this round in order to improve your evaluation. One of the 15 rounds will be randomly chosen to determine the payoffs at the end of the experiment. This means that every round could be the payoff-relevant round. The decisions that were made in the randomly chosen round determine the payoffs for both participants.

(iii) The third part of your payoff depends on your expectations. One of the 15 rounds will be randomly chosen to determine the payoffs at the end of the experiment. The round that was selected for the second part of your payoff will never be selected. Except for the round that was selected for the second part of your payoff, each round may be payoff relevant. You can earn an additional 4 euros every time your expectations regarding your partner's decision turn out to be correct. Otherwise, your payout is zero. You only earn the additional 4 euros if your partner chooses the decision you expected.

After the last round is completed there will be a brief questionnaire. Afterwards, you will receive your payoffs in cash.

The experiment will begin shortly. If you have any questions please raise your hand and wait calmly until someone comes to you. Please do not talk to the other participants at any time during the experiment. Thank you for participating.