Journal of Behavioral and Experimental Economics 83 (2019) 101480

Contents lists available at ScienceDirect

Journal of Behavioral and Experimental Economics

journal homepage: www.elsevier.com/locate/jbee

Reciprocity through ratings: An experimental study of bias in evaluations☆

Simon D. Halliday*,[a], Jonathan Lafky[b]

[a] Smith College Department of Economics, Pierce Hall 107, 21 West St, Northampton, MA 01062, USA
[b] Carleton College Department of Economics, Northfield, MN 55057, USA

ARTICLE INFO

Keywords:
Ratings
Reciprocity
Punishment

JEL classification:
D47
D82
D83

ABSTRACT

This paper studies the potential for ratings of seller quality to be influenced by side payments to raters. In a laboratory setting, we find that even modest side payments from sellers to raters have large effects, with the type of rating (favorable or unfavorable) given to a seller determined primarily by how large a monetary transfer the seller makes to the rater. Our results demonstrate that side payments can crowd out a rater's concern for buyers, even in situations where there is no potential for long-term relationship building.

1. Introduction

Online commerce is a major component of the economy, accounting for $389 billion in sales in 2016, or 8.0% of all US retail.[1] This is an increase of 14.4% from 2015, compared to only 2.8% growth in total retail. Rating systems are one of the main facilitators of online (and many offline) transactions, allowing consumers to share information about a product's quality with future buyers. The usefulness of that information is not guaranteed, however, as many rating systems do not provide any explicit reward for giving accurate ratings, leaving open the possibility that other, potentially harmful, incentives may influence the decision to rate. This paper experimentally studies the effect of one such incentive: side payments made by sellers in an attempt to influence ratings. We show that such payments can heavily influence rater behavior, decreasing the usefulness of ratings for future consumers.

In many settings, consumer-generated ratings are public goods and are regularly inefficiently under-supplied in free markets (Avery, Resnick, & Zeckhauser, 1999). As a result, encouraging the provision of online ratings has been an emphasis in existing ratings research, leading to several explanations as to why consumers provide ratings, and how institutions can encourage more frequent rating (Chen, Harper, Konstan, & Li, 2010; Li & Xiao, 2014). The main challenge when attempting to encourage ratings comes from the need to overcome small, often implicit, costs that discourage consumers from rating, while also maintaining the accuracy of any ratings that are generated. Some incentives, whether external rewards, or internal motivations, can bias ratings sufficiently to diminish or even eliminate all benefits to consumers (Lafky & Wilson, 2015).

One motivation to rate is the desire to reciprocate good treatment with good ratings (positive reciprocity) and ill treatment with poor ratings (negative reciprocity). Both negative and positive reciprocity have been well-documented, showing that individuals are willing to forgo earnings to punish or reward behavior based on perceived fairness (Fehr & Gächter, 2002; Fehr & Fischbacher, 2004; Nikiforakis & Mitchell, 2014; Balafoutas, Nikiforakis, & Rockenbach, 2014; Nosenzo, Offerman, Sefton, & van der Veen, 2015).[2] Reciprocity has been demonstrated to play a role in rating behavior in the field, in settings such as Airbnb (Fradkin, Grewal, & Holtz, 2019) and eBay (Resnick & Zeckhauser, 2002). The role of reciprocity is especially visible in two-sided rating environments, in which both buyers and sellers rate one another. When both sides of the market can rate one another, ratings are used strategically to punish or reward other members of the market

☆ This work was supported by research funds from Smith College, Lafayette College and Carleton College.
* Corresponding author.
E-mail addresses: shalliday@smith.edu (S.D. Halliday), jlafky@carleton.edu (J. Lafky).
[1] U.S. Census Bureau, U.S. Retail Trade Sales - Total and E-commerce (1998–2016). Retrieved from www.census.gov/data/tables/2016/econ/arts/annual-report.html
[2] Furthermore, and importantly for our treatments, willingness to punish has also been shown to follow the law of demand: it decreases in incidence as the price of punishment increases in public goods games (Anderson & Putterman, 2006), third-party punishment games (Carpenter, 2007) and in reviewing contexts (Lafky, 2014).

https://doi.org/10.1016/j.socec.2019.101480
Received 28 May 2019; Received in revised form 18 August 2019; Accepted 16 October 2019
Available online 18 October 2019

and to protect one's own reputation (Klein, Lambertz, Spagnolo, & Stahl, 2006; Klein, Lambertz, Spagnalo, & Stahl, 2009; Bolton, Greiner, & Ockenfels, 2013). Such strategic concerns can damage the informational content of ratings, as raters become more concerned with protecting their own reputation than with accurately describing quality.

One-sided rating environments, which are the focus of this paper, minimize such strategic considerations. Raters in one-sided systems have been shown to exhibit altruistic concern for informing buyers, as well as reciprocity toward sellers based on the seller's choice of quality (Lafky, 2014). Reciprocity based on seller quality is not, in itself, troubling as raters face no dilemma over whom to favor with their ratings. If a rater gives an unfavorable rating intended to punish a seller for offering low quality or a favorable rating intended to reward a seller for offering high quality, they are also informing other buyers of the seller's true quality.

A natural, though potentially harmful, way for sellers to exploit reciprocity is through the use of side-payments to raters. Side-payments and gifts have been experimentally investigated in a variety of economic contexts, primarily focused on public sector bribery and corruption, harassment bribes (Abbink, Dasgupta, Gangadharan, & Jain, 2014), or petty bribes between officials and citizens (Barr & Serra, 2010). Similarly, gift-giving has been shown to corrupt principal-agent relationships, with principals more likely to act against an agent's interests following a gift from an interested third-party (Malmendier & Schmidt, 2017). A rater who desires to reciprocate a seller's side-payment (or lack thereof) may be tempted to do so at the expense of accurate ratings, resulting in two competing objectives: reciprocity toward the seller in response to a side-payment, and a desire to accurately inform future buyers.

At their simplest, side-payments may take the form of a direct monetary bribe from a seller to a rater. In practice, side-payments might be less explicit, embodied by more subtle forms of generosity. For example, a company might send an evaluation copy of a product to a reviewer, allowing the reviewer to keep the item after the review is complete as a subtle form of gift.[3] We ask whether such favorable treatment, separate from the product's underlying quality, will influence the type of rating chosen for the company. Specifically, we consider environments in which a payment toward the rater precedes the choice of rating and cannot be contingent upon the rating given.

When side-payments are possible, it becomes difficult to disentangle whether ratings are motivated by reciprocity over quality versus reciprocity over side-payments. To isolate reciprocity over side-payments, we implement a simple, stylized experiment in which the seller's quality is *exogenously* determined. In other words, we strip away the seller's quality decision, and therefore responsibility for the quality that they offer, leaving them with only the choice of side-payment. By exogenizing quality, we remove the effect of reciprocity over product quality, and are able instead to cleanly observe the relative strength of a rater's reciprocity toward sellers versus their desire to aid future buyers. Because sellers are not responsible for their quality, we do not have our raters "buy" from the seller, meaning that their payoff is not directly affected by the seller's quality. While raters in the field are often themselves buyers who have experienced the seller's quality, this design again helps to isolate the tension raters face between concern for transfer size and concern for future buyers.

As most rating environments exhibit at least a small cost of rating, arising from the opportunity cost of a rater's time, or the nominal mental effort required to evaluate a product (Kamei & Putterman, 2018), we examine the effect of side payments in two experimental environments, one in which rating is costly and one in which it is free. While the costly treatment provides a more realistic description of typical rating systems in the field, the free treatment allows us to cleanly observe the tradeoff raters face between concern for buyers and concern for sellers. Although we are primarily interested in the effect of side-payments on ratings, the difference in costs between treatments also allows us to examine the opposite direction, asking whether the ease with which consumers can punish or reward via ratings influences how generous sellers are with side-payments. This is similar to Xiao (2013) who studies the effectiveness of punishment when the decision to punish is either free or actually beneficial to the punisher.

We are not the first to suggest that sellers may manipulate ratings in their favor, for example by fraudulently reviewing themselves or their competitors (Dellarocas, 2006; Mayzlin, Dover, & Chevalier, 2014), or by exploiting a good reputation to sell low-quality products (Dini & Spagnolo, 2009). Our contribution is to use a controlled environment to show that non-contingent side-payments, even without the possibility of future interactions, can heavily influence ratings.

Our results show that the size of the transfer from the seller is the primary determinant of the type of rating sent. That is, simply giving a gift is not sufficient to curry favor with a rater, but the magnitude of the gift determines the likelihood of a favorable rating. Raters give favorable ratings when sellers make large transfers and unfavorable ratings for small transfers, regardless of the seller's underlying quality. As a result, we observe many "misleading" ratings that are not in buyer's best-interest. We also find that sellers are sensitive to the cost of rating, giving smaller transfers when it is more costly for consumers to provide ratings.

Before discussing our results, we first develop of simple model of rating behavior, initially for strictly-self interested agents, and then when raters have social preferences.

2. Theory

Our environment consists of three players, a seller, a rater and a buyer. The game begins with the seller drawing a quality q from the uniform distribution $U[0, q_{\max}]$, and choosing an amount of money, $a \in [0, a_{\max}]$, to transfer to the seller. The seller's quality is exogenously determined in order to isolate rater reciprocity over the size of the transfer, separate from possible reciprocity over the seller's quality. The rater learns the seller's q and a, and then has the option to pay a cost $c \geq 0$ in order to send a binary rating $r \in \{G, B\}$ (i.e., Good or Bad), or pay no cost and send the empty message $r = \varnothing$.

The buyer next learns what rating, if any, was given, and must choose either the seller or an unknown outside option. If the buyer chooses the seller, the buyer receives a payoff equal to q, while the seller receives a fixed payoff π that does not depend upon q. If the buyer chooses the outside option, the buyer's payoff is a random draw, k, from $U[0, q_{\max}]$, and the seller's payoff is zero. The distribution of the outside option is the same as that of q, meaning that the outside option can be effectively thought of as an alternative untried seller for the buyer to choose. Denote the utility of the seller, rater, and buyer by U_S, U_R, and U_B, respectively.

If we assume strictly self-interested preferences, the prediction in this environment is straightforward. Raters exhibit no reciprocity based on transfer size, which means that sellers have no ability to influence ratings and therefore always choose transfers of zero. Similarly, raters have no incentive to inform buyers of seller quality. In short, sellers make no transfers, raters do not rate, and buyers receive no information as to whether the seller or outside option is a better choice.

Next, consider the model in which the rater has other-regarding preferences for both the seller and the buyer, given by

[3] We refer specifically to instances in which free products are a form of preferential treatment for reviewers over general customers, such as in the case of the Amazon Vine program. Such exchanges are distinct from, for example, including a bonus item free of charge with every purchase. A company that routinely bundles a product for free to *all* customers (e.g., free headphones with purchase of a new phone) is essentially just selling a higher-quality product, while reviewers who preferentially receive free demo products have a more favorable experience than can reasonably be expected for the general customer.

$$U_R = \alpha(a)\cdot U_S + \beta\cdot U_B - c\cdot \mathbb{1}_{rate}$$

where $\alpha(a)$ and β measure the rater's concern for the seller and buyer, respectively, and $\mathbb{1}_{rate}$ is an indicator function for whether a rating was given. We assume that there is some a^* such that $\alpha(a) > 0$ when $a > a^*$ and $\alpha(a) < 0$ when $a < a^*$, but we do not assume any specific functional form.[4] The rater therefore exhibits either positive or negative concern for the seller, depending upon how generous the seller was with the choice of a. If the seller was (subjectively) generous, then the rater exhibits altruistic concern for the seller, however if the seller is perceived as being selfish, then the rater prefers to see the seller worse off instead. Unlike $\alpha(a)$, the rater's concern for the buyer, β, is not a function but a fixed parameter. This is because the rater does not observe any behavior from the buyer on which to condition positive or negative concern. The rater's payoff is not directly influenced by the seller's exogenously determined quality, and as a result the rater's desire for reciprocity depends only on the transfer, a, and not on the quality, q.

The rater will give a positive rating, $r = G$, if the utility from rating is greater than the utility from not rating:[5]

$$\alpha(a)\cdot U_{S|r=G} + \beta\cdot U_{B|r=G} - c > \alpha(a)\cdot U_{S|r=\varnothing} + \beta\cdot U_{B|r=\varnothing}$$

For simplicity, we assume that the buyer treats ratings credulously and randomizes over the choices if no rating is sent.[6] We next substitute for U_S and U_B to get:

$$\alpha(a)\cdot\pi + \beta\cdot q - c > \alpha(a)\cdot\frac{\pi}{2} + \beta\cdot\frac{q + E(k)}{2}$$

which can be rewritten as:

$$\alpha(a)\cdot\pi + \beta(q - E(k)) > 2c \tag{1}$$

Intuitively, Eq. (1) says that the rater is more likely to give a positive rating the more generous the seller is (larger a), the higher the seller's quality (larger q) or the lower the cost of rating (smaller c). Generosity and quality substitute for one another, so a high-quality seller can get away with being less generous to the rater, whereas a low-quality seller must be more generous to receive a positive rating.

Likewise, the rater will give an unfavorable rating, $r = B$, if:

$$\alpha(a)\cdot 0 + \beta\cdot E(k) - c > \alpha(a)\cdot\frac{\pi}{2} + \beta\cdot\frac{q + E(k)}{2}$$

or equivalently:

$$\alpha(a)\cdot\pi + \beta(q - E(k)) < -2c \tag{2}$$

The intuition conveyed by Eq. (2) is similar to before: the rater is more inclined to give a negative rating the less generous the seller is, the lower the seller's quality, or the lower the cost of rating.

When neither of Eqs. (1) and (2) hold, the rater prefers to give no rating. In other words, the rater gives no rating when both

$$\alpha(a)\cdot\pi + \beta(q - E(k)) < 2c$$

and

$$\alpha(a)\cdot\pi + \beta(q - E(k)) > -2c$$

hold simultaneously. Overall, our model predicts that as the sum of the utilities the rater experiences due to social preferences increases, the rater transitions first from giving a negative rating, to no rating, to eventually giving a positive rating.

We summarize our analysis with four predictions, which we later take to the lab:

Prediction 1. Raters may rate even when it is costly to do so.

Eqs. (1) and (2) show that both conditional altruism toward the seller and unconditional altruism toward the buyer will motivate raters to pay the cost of rating.

Prediction 2. High-quality sellers make smaller transfers than low-quality sellers.

Due to the tradeoff between quality and transfer size in Eq. (1), lower quality sellers need to offer larger transfers than their high-quality equivalents in order to generate the same rating.

Prediction 3. Choice of rating depends upon both quality and the size of the transfer. A rating is more likely to be favorable (unfavorable) when quality is higher (lower) or when a larger (smaller) transfer occurs.

As quality or transfer become more extreme (either high or low), the greater is the rater's concern due to social preferences. Sellers of moderate quality or transfers may not elicit a sufficiently strong level of concern from the rater to overcome the cost of rating.

Prediction 4. Raters will mislead buyers to reward or punish sellers for the size of their transfers.

If reciprocity toward sellers dominates altruism toward buyers, ratings will be given solely to aid sellers and will not reveal actual seller quality. We next evaluate these predictions against experimental data.

3. Experimental design and procedures

Our experiment was conducted at the Cleve E. Willis Experimental Economics Laboratory at the University of Massachusetts, Amherst using z-Tree experimental software (Fischbacher, 2007). A total of 396 subjects were recruited from the student population using the ORSEE recruitment system (Greiner, 2015).

Each session lasted 45 minutes on average, including time for instructions, participation in the experiment, a demographic questionnaire and being paid in private at the conclusion of the session. Subjects were randomly assigned to cubicles that were separated from each other visually and physically, and subjects were prohibited from speaking. All subjects received a common set of instructions and all questions were answered privately.

We implemented a one-shot between-subject design with two treatments.[7] Participants were randomly assigned to groups of three

[4] Different models of reciprocity compatible with our environment include Rabin (1993); Falk and Fischbacher (2006); Charness and Rabin (2002). Our concern is only with the qualitative predictions of the reciprocity function, that raters exhibit positive reciprocity for sufficiently high transfers, and negative reciprocity for sufficiently small transfers.

[5] In order to give a positive rating, the rater actually needs to prefer $r = G$ to both $r = \varnothing$ and $r = B$, however the condition for preferring $r = G$ to $R = \varnothing$ implies the condition for preferring $r = G$ to $r = B$. Similarly, when we consider the decision to give a negative rating, below, we need only compare $r = B$ to $r = \varnothing$. These comparisons are made explicit in the appendix.

[6] For the buyer to behave credulously and follow the rater's advice, we only need that $E(q|r = G) > E(q)$, meaning that positively rated sellers have a higher expected value than the outside option, and $E(q|r = B) < E(q)$, meaning that negatively rated sellers have a lower expected value than the outside option. In other words, we require only that ratings indicate the better option more often than not. If these conditions do not hold, the buyer either ignores ratings entirely, or infers the opposite of a rating's literal meaning (e.g., choose the seller with a negative rating, avoid the seller with a positive rating). In both instances we end up with a babbling equilibrium in which no communication occurs and no transfers are made. Babbling is always possible in communication environments, but as we will show later, we do not see evidence for such behavior in our experiment, as buyers overwhelmingly follow rater recommendations.

[7] A variety of experiments with punishment, deceit and third parties use one-shot games to remove the reputational concerns of repetition and to increase the salience of the single choice. Fehr and Fischbacher (2004) and Carpenter (2007) use one-shot choices for third-party punishment. Falk and Kosfeld (2006), Ziegelmeyer, Schmelz, and Ploner (2012) and Burdìn, Halliday,

people. Within each group, participants were randomly assigned the roles of seller, rater and buyer, which we referred to in the experiment as Participant A, Participant B and Participant C, respectively. Every group contained exactly one participant of each type.

Within each group, the seller was randomly assigned an integer drawn from 1 to 16 inclusive.[8] This draw represented seller quality, though it was referred to in the experiment simply as "the X number." Randomly drawing quality provides us with crucial exogenous variation to examine the relationship between quality, transfers and choice of rating. After learning the randomly chosen quality, the seller was given an endowment of \$16, and was given the option to transfer any amount of that endowment to the rater in \$1 increments.

The rater learned the amount of money that the seller chose to transfer, as well as the quality that was randomly chosen for the seller. The rater chose either to pay a cost, $c \geq 0$, to send a message to the buyer, or to not incur a cost and send no message. The rater had a choice of two messages, either "Choose Participant A" or "Do not choose Participant A." In the discussion below we refer to a message recommending the seller as a "positive rating" and one recommending against the seller as a "negative rating," though we did not use these terms in the experimental instructions. The cost the rater incurred depended on the treatment as follows:

$$c = \begin{cases} \$0 \text{ in the free treatment} \\ \$0.50 \text{ in the costly treatment} \end{cases}$$

The buyer learned what message the rater sent, if any, and then chose between the seller and an outside option. In the experiment we referred to the outside option as "the Y number," which was another integer drawn randomly from 1 to 16. The buyer earned \$1 times the seller's quality if they chose the seller, and \$1 times the outside option otherwise. As before, we denote the seller transfer by a, which gives us two earnings profiles for each participant:

$$\text{Seller} \quad \pi_S = \begin{cases} \$32 - a & \text{if chosen by buyer} \\ \$16 - a & \text{if not chosen by buyer} \end{cases}$$

$$\text{Rater} \quad \pi_R = \begin{cases} a - c & \text{if rater chose to send a message} \\ a & \text{if rater chose not to send a message} \end{cases}$$

$$\text{Buyer} \quad \pi_B = \begin{cases} X & \text{if buyer chose seller} \\ Y & \text{if buyer chose outside option} \end{cases}$$

In the free treatment, $c = 0$, and therefore rater payoffs collapse to simply $\pi_R = a$ for both sending and not sending a message.

Subject understanding of choice sets and payoffs was assured by four quiz questions at the end of the instructions, followed by three control questions in the experimental software. Once all subjects had answered the quiz and control questions correctly (with opportunities to ask questions privately), subjects made their choices in the sequence described above: first seller, then rater, then buyer. After making their experimental choices, each subject completed a basic demographic survey. The survey was not incentivized and subjects were told that their responses on the survey were not connected to their final payment. After completing the survey, subjects were called to a cubicle in order to privately receive their earnings from the experiment, plus a \$5.00 show-up fee.

(*footnote continued*)
and Landini (2018) use a one-shot game in two and three-party principal-agent (and monitor) games. Gneezy (2005) and Sutter (2009) use a one-shot sender-receiver game to understand deception and communication.

[8] The instructions indicated that this number and the outside option described below were drawn from 0 to 16 inclusive, but due to a programming error both were actually drawn between 1 and 16 instead. This error was not detected until after the sessions were completed. The effect on subject behavior should have been minimal, and subjects were not harmed by this change, as it meant that mean quality and therefore payoffs were slightly higher than intended.

4. Results

Table 1 shows summary statistics for the experiment. There were no statistically significant differences by treatment or across experimental session in demographic characteristics. Due to the lower frequency of rating and resultant decrease in the number of ratings observed, a larger sample size was necessary in the costly treatment relative to the baseline free rating treatment in order to observe a similar number of ratings in each treatment. We therefore ran twice as many sessions in the costly treatment (12 sessions) as in the free treatment (6 sessions). The subjects' earnings do not substantially differ across treatments, with average earnings (net of the show-up fee) of approximately \$11. Raters earn the least on average and sellers earn the most, consistent with the design of the experiment.

Fig. 1 shows seller decisions made in each treatment. Each dot represents the seller's quality on the horizontal axis and the proportion of the seller's \$16 endowment that was transferred on the vertical axis. As Table 1 shows, a large majority (86%) of sellers transfer positive amounts, and there is no significant difference in the frequency of non-zero transfers between the treatments, with 89% sharing at least \$1.00 in the free treatment and 85% doing so in the costly treatment. Consistent with Xiao (2013), sellers are responsive to the likelihood of punishment from raters, sending average transfers of \$5.17 in the costly treatment and \$6.78 in the free treatment, an increase of 31.14% when rating is free (Mann–Whitney $z = 2.67$, $p = 0.008$).

Table 2 shows the determinants of transfer size from Tobit (odd columns) and OLS (even columns) regressions. As before, sellers share a smaller portion of their endowment when it is costly for raters to give a rating. Transfer size also depends upon the seller's quality, with higher-quality sellers choosing larger transfers, though this effect is relatively modest, with a 1-unit increase in quality resulting in a \$0.17 increase in transfer size. This behavior is somewhat puzzling, and is opposite from our theoretical prediction that quality and transfer size serve as substitutes. Note that quality is no longer statistically significant in specifications (3) and (4) in Table 2. Restricting our regressions by treatment shows a statistically significant and positive relationship between quality and transfer size in the costly treatment, but no such correlation in the free treatment. In other words, in the costly treatment, high-quality sellers give slightly larger transfers than low-quality sellers, while in the free treatment transfer size does not vary by seller quality. We report the results of the restricted regressions in Table A1 the appendix. Seller behavior in the free treatment can be explained by sellers who expect raters to have very little concern for buyers, which, as we discuss below, appears to be an empirically sound belief. Seller behavior in the costly treatment is more difficult to explain. A belief that raters are concerned only with transfers should result in sellers being insensitive to their own quality, while a belief that raters are concerned with quality should result in sellers who treat quality and transfer size as substitutes, as explained in the theory section above. Finally, a seller who believes raters to be concerned only with quality should give no transfer. We revisit the question of seller behavior in the conclusion, and we control for quality and transfer size in our analyses of rating behavior below.

Our primary interest is in how raters respond after learning seller quality and choice of transfer. We observe in Table 1 that the choice to rate is highly sensitive to costs. More than twice as many raters (84.44%) choose to rate in the free treatment as in the costly treatment (41.38%) (Mann–Whitney $z = 4.708$, $p < 0.001$). The increase in ratings moving from costly to free is perhaps less impressive than the fact that almost half of all raters choose to rate in the costly treatment. A large number of raters are still motivated to rate even when it is costly to do so.

Result 1. Many raters (41.38%) rate even when it is costly to do so.

We expand on Result 1 through a series of regressions on the likelihood that a rater will send *any* rating, with a binary dependent

Table 1
Summary statistics.

	Free	Costly	Pooled
Demographics			
Sessions	6	12	18
Groups	45	87	132
Subjects	135	261	396
% Female	44.44	41.00	42.17
Age	20.42	20.45	20.44
	(1.41)	(1.94)	(1.78)
Sellers			
Proportion transferred	0.42	0.32	0.36
	(0.21)	(0.22)	(0.22)
Transfer in $	6.78	5.17	5.72
	(3.36)	(3.48)	(3.51)
Proportion of transfers > 0	0.89	0.85	0.86
	(0.32)	(0.36)	(0.34)
Proportion transferred $\|r = G$	0.52	0.46	0.49
	(0.18)	(0.21)	(0.20)
Proportion transferred $\|r = B$	0.32	0.20	0.26
	(0.23)	(0.18)	(0.21)
Raters			
Prob. rating sent $(r \neq \emptyset)$	0.84	0.41	0.56
	(0.37)	(0.50)	(0.49)
Prob. $r = G$ $\|r \neq \emptyset$	0.50	0.50	0.50
	(0.51)	(0.51)	(0.50)
Prob. misleading rating $\|r \neq \emptyset$	0.39	0.26	0.33
	(0.50)	(0.44)	(0.47)
Buyers			
Prob. chose seller	0.51	0.46	0.48
	(0.51)	(0.50)	(0.50)
Prob. chose seller $\|r = G$	0.95	0.83	0.89
	(0.23)	(0.38)	(0.32)
Prob. chose seller $\|r = B$	0.05	0.06	0.05
	(0.23)	(0.24)	(0.23)
Earnings			
Seller earnings	17.40	18.18	17.92
	(7.74)	(8.24)	(8.05)
Rater earnings	6.78	4.97	5.58
	(3.36)	(3.49)	(3.54)
Buyer earnings	9.27	8.23	8.58
	(4.92)	(4.21)	(4.48)
All subject earnings	11.15	10.46	10.69
	(7.22)	(8.00)	(7.74)

Earnings are in dollars, excluding the show-up fee. Standard deviations in parentheses.

variable that takes the value 1 if the rater sends a rating and 0 if not. Regression results are shown in Table 3, with probit specifications in the odd columns and OLS in the even columns. In each of the specifications there is a large negative effect from the cost of rating and the size of this effect increases as we add more controls, suggesting that the choice to rate is heavily influenced by the presence of a cost. Consistent with Result 1, the regression results show a rater's likelihood of rating drops between 41 and 65 percentage points when rating is costly. In specifications 5 and 6 we also check the predicted tendency to rate more frequently when quality or transfers tend toward high or low levels, and less frequently for moderate levels. We see modest evidence of this behavior, with small positive correlations and mixed statistical significance.

The size of transfer does not influence the likelihood that a rater rates, however once we condition upon a rating being sent we find that transfers are highly influential. Negative ratings were preceded by average transfers of $4.14, compared to $7.84 for positive ratings (Mann–Whitney $z = -4.26$, $p < 0.001$). For a given rating, the amounts transferred are not statistically significantly different across treatments. Table 4 expands on this simple comparison through a series of regressions examining the choice of rating, conditional on any rating being sent. These regressions include seller quality, which controls for the weak relationship between quality and transfer size, allowing us to focus on the effect of transfer size on choice of rating. We model the choice between the two possible ratings as a binary variable that takes the value 1 for a positive rating and 0 for a negative rating. In every specification we find a statistically significant relationship between transfer size and choice of rating, with each additional dollar transferred leading to approximately a 6 percentage point increase in the likelihood that the rating is favorable.[9] As a robustness check, Tables A3 and A4 in the appendix report the results of ordered probit regressions that combines the choice of rating with the decision to rate at all. The categories are a negative rating (0), no rating (1), and a positive rating (2). The results remain substantively the same: the amount received by a rater corresponds to a lower probability of negative rating and a higher probability of a positive rating.

Result 2. A rater's choice of rating is determined by the size of transfer from the seller.

Result 2 is our main finding, as it indicates that the type of rating sent is determined entirely by the seller's choice of side-payment, and not by seller quality. The result suggests that side-payments can fully crowd out the rater's desire to accurately describe quality. This behavior is important and possibly troubling, as it suggests that raters' intrinsic motivation to help other consumers may be small, relative to other factors. Even with no future rewards at stake, raters are more concerned with reciprocating seller side-payments than with accurately informing buyers.

The strong relationship between transfer size and choice of rating produces many ratings that are not in a buyer's best interest, and a rater's choice to punish or reward a seller often comes at the buyer's expense. We say that a rater misleads a buyer if they provide a favorable rating when quality is less than 8 (an overstatement) or an unfavorable rating when quality is greater than 8 (an understatement). Note that misleading ratings are not, strictly speaking, dishonest, as raters are only making recommendations, and neither of the messages they can send (e.g., "Choose Participant A") is ever technically false. Across both treatments 32.9% of ratings are misleading, with more misleading ratings in the free treatment (39.5%), than the costly (25.7%), though that difference is not statistically significant.

Result 3. Nearly one-third (32.9%) of all ratings are misleading.

Examples of misleading ratings can be seen in Fig. 1, where the dashed vertical line represents the ex-ante expected quality level. Any favorable rating to the left of the vertical line overstates quality, instructing the buyer to choose the seller even though the seller is of below-average quality. Similarly, unfavorable ratings to the right of the vertical line understate quality, instructing the buyer to avoid the seller, despite the seller having above-average quality.

Although we observe that approximately one-third of ratings are misleading, those numbers likely under-represent raters' actual willingness to mislead buyers. We would not expect to see more than half of all ratings be misleading, even if raters were solely motivated by concern for sellers. For example, the act of punishing a seller for a small transfer only results in misleading behavior when the seller is of high quality, as the remaining half of the time the seller is of low quality and an unfavorable rating intended to punish coincidentally helps the buyer to choose their best option. In other words, the actual proportion of the population willing to give misleading ratings is approximately double what we observe, suggesting that many raters are willing to make buyers worse off in order to punish or reward sellers for their side-payments.

It is possible that risk aversion could cause a rater motivated solely by altruism toward buyers to give a favorable rating to a below-average seller. That said, observed rates of misleading ratings are very similar if

[9] This result is robust to instead using dummy variables for each quartile of amount received or a dummy variable (with interaction term with cost) for whether the amount received was above or below the median transfer of $6.

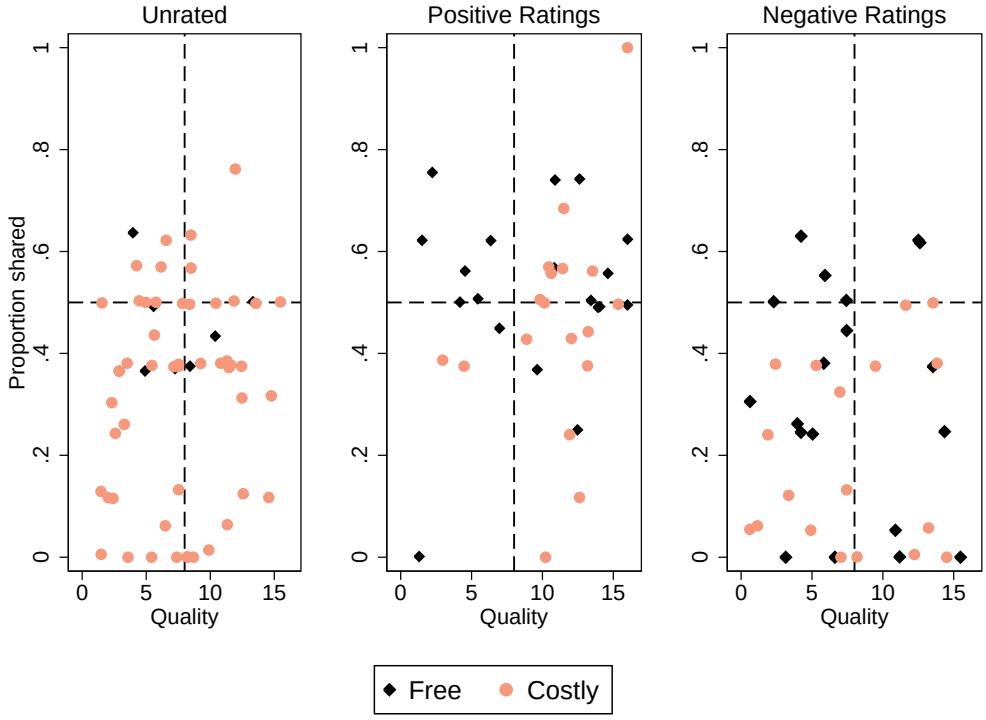

Fig. 1. The three panels show the quality and transfer for sellers who received no rating, a positive rating, or a negative rating, respectively. A small amount of noise has been added to show overlapping points.

Table 2
Amount transferred by seller.

	(1) Tobit	(2) OLS	(3) Tobit	(4) OLS
Cost	−1.646**	−1.579**	−2.886*	−2.773**
	(0.688)	(0.607)	(1.516)	(1.316)
Quality	0.167**	0.156**	0.0752	0.0672
	(0.0794)	(0.0701)	(0.132)	(0.115)
Female	−0.889	−0.901	−0.835	−0.850
	(0.680)	(0.595)	(0.679)	(0.596)
Age	−0.169	−0.142	−0.170	−0.144
	(0.189)	(0.164)	(0.184)	(0.160)
Quality × Cost			0.148	0.143
			(0.163)	(0.144)
Constant		8.740**		9.516***
		(3.392)		(3.279)
N	132	132	132	132
Adjusted R^2		0.072		0.072

Robust standard errors in parentheses. 18 observations left-censored at 0; 1 observation right-censored at 16. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

we change the threshold used to define a misleading rating, either upward to 9 or downward to 7. Additionally, if risk aversion did lead to favorable ratings for below-average sellers, we would expect to see more overstatements than understatements, as there is no similar incentive to negatively rate above-average sellers. This is not the pattern we observe in the data, with no significant different in the frequency of understatements and overstatements, at 31.25% and 34.15%, respectively ($p = 0.80$, Mann–Whitney).

Given our experimental parameterization, the seller's reward from being chosen ($16.00) is greater or equal to the maximum difference in earnings to the buyer from choosing between the seller and the outside option. As a result, if a rater were solely concerned with efficiency (i.e., the sum of the seller's and buyer's payoffs) they would give favorable ratings to all sellers, regardless of quality, resulting in many overstatements. We do not observe such a pattern in the data, as overstatements and understatements are equally common.

Despite the frequency with which ratings are misleading, when a rating is sent, buyers generally follow the recommendation made by the rater. The seller is chosen following a favorable rating 94.7% of the time in the free treatment and 83.3% of the time in the costly treatment (see Table 1). Following a negative rating, 5.3% and 5.6% of sellers are chosen in the free and costly treatments, respectively. If no rating is sent, buyers appear to randomize over choosing the seller and the outside option, with 48.3% of buyers in the free treatment and 47.3% of buyers in the costly treatment choosing the seller.[10] None of the buyer behavior is significantly different between treatments. We summarize these findings with our next result.

Result 4. Buyers generally follow rater recommendations.

Buyers generally trust ratings, but what effect do ratings have on buyers? We first examine ex-ante buyer welfare based on both the empirical frequency with which each type of rating is given, and the expected quality conditional upon each type of rating. In other words, within each treatment the expected payoff to a potential buyer is given by:

$$\begin{aligned}\pi_{B,T} &= P_T(\text{No rating}) \times (0.5E[X|\text{No rating}] + 0.5E[Y]) \\ &\quad + P_T(\text{Positive rating}) \times E_T[X|\text{Positive rating}] \\ &\quad + P_T(\text{Negative rating}) \times E[Y] \end{aligned} \tag{3}$$

Where $\pi_{B,T}$ is the expected payoff to a credulous buyer in treatment T, P_T is an empirical probability, $E_T[X|\text{Positive Rating}]$ and $E[X|\text{No rating}]$ are the mean quality following a favorable rating or no rating, respectively, and $E[Y]$ is the unconditional expected value of Y, which does not vary by treatment. We find that $\pi_{B,C} = 8.74$ in the costly treatment, and $\pi_{B,F} = 8.78$ in the free treatment, meaning that in both treatments the ex-ante buyer payoffs are very close to the a seller's ex-

[10] In regressions reported in Table A2 of the appendix, we show that the unconditional probability of the buyer choosing the seller does not correlate with the treatment or receiving a rating of any quality (good or bad). On the contrary, conditional on receiving a rating, a positive rating increased the probability the buyer will choose the seller by approximately 87%.

Table 3
Decision to send a rating (1 = sent, 0 = not sent).

	(1) Probit	(2) OLS	(3) Probit	(4) OLS	(5) Probit	(6) OLS
Cost	−0.411***	−0.426***	−0.589***	−0.625***	−0.646***	−0.643***
	(0.0666)	(0.0782)	(0.172)	(0.168)	(0.177)	(0.156)
Quality	0.0144*	0.0142	0.00382	0.00203	0.000568	−0.00184
	(0.00875)	(0.00884)	(0.0129)	(0.00905)	(0.0135)	(0.00819)
Transfer	−0.00367	−0.00308	−0.0121	−0.00784	0.00458	0.00958
	(0.0105)	(0.0108)	(0.0148)	(0.0100)	(0.0199)	(0.0128)
Female	−0.123	−0.131	−0.126*	−0.133	−0.123*	−0.138*
	(0.0756)	(0.0807)	(0.0756)	(0.0816)	(0.0742)	(0.0819)
Age	0.00374	0.00367	0.00306	0.00405	0.00892	0.00799
	(0.0241)	(0.0243)	(0.0239)	(0.0238)	(0.0246)	(0.0262)
Quality × Cost			0.0143	0.0197	0.0156	0.0219
			(0.0171)	(0.0161)	(0.0172)	(0.0158)
Transfer × Cost			0.00978	0.00531	0.0203	0.00865
			(0.0197)	(0.0183)	(0.0225)	(0.0168)
|Quality − 8|					0.0356**	0.0351*
					(0.0171)	(0.0183)
|Transfer − 8|					0.0400	0.0326*
					(0.0256)	(0.0185)
Constant		0.721		0.848*		0.457
		(0.510)		(0.490)		(0.550)
N	132	132	132	132	132	132
Adjusted R^2		0.177		0.171		0.200

Robust standard errors in parentheses. Probit coefficients are marginal effects. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 4
Type of rating chosen (1 = positive, 0 = negative).

	(1) Probit	(2) OLS	(3) Probit	(4) OLS
Cost	0.0824	0.0806	−0.157	−0.0308
	(0.101)	(0.104)	(0.311)	(0.277)
Quality	0.0151	0.0160	0.00837	0.0115
	(0.0105)	(0.0113)	(0.0135)	(0.0155)
Transfer	0.0630***	0.0668***	0.0554***	0.0646***
	(0.0109)	(0.0135)	(0.0191)	(0.0208)
Female	−0.0272	−0.0469	−0.0308	−0.0485
	(0.0982)	(0.101)	(0.0979)	(0.104)
Age	0.0235	0.0256	0.0265	0.0253
	(0.0381)	(0.0389)	(0.0380)	(0.0398)
Quality × Cost			0.0165	0.0108
			(0.0227)	(0.0238)
Transfer × Cost			0.0146	0.00203
			(0.0333)	(0.0291)
Constant		−0.587		−0.527
		(0.847)		(0.927)
N	74	74	74	74
Adjusted R^2		0.236		0.216

Robust standard errors in parentheses. Probit coefficients are marginal effects. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 5
Buyer outcomes.

	Helpful Rating	Misleading Rating	Difference
Costly treatment	9.85	6.89	2.96*
	(0.76)	(1.45)	
N	27	9	
Free treatment	10.74	7.33	3.41**
	(1.06)	(1.21)	
N	23	15	
Combined sample	10.26	7.17	3.09***
	(0.63)	(0.91)	
N	50	24	

Standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$, via Mann–Whitney.

ante expected quality of 8.5. This shows that, on average, buyers are just as well-off as if they were choosing between their options at random, which is consistent with our findings that ratings are shaped entirely by the size of the seller transfer.

Although ratings do not influence buyer outcomes on average, individual ratings can be helpful or harmful to buyers. Table 5 reports final buyer outcomes following helpful (positive ratings for above average sellers or negative ratings for below average sellers) and misleading ratings in each treatment, showing the change in welfare that result from each type of rating. Buyers are consistently better off when ratings helpfully recommend the higher-payoff option than when they are misleading, regardless of the cost of rating. Switching from a misleading rating to a helpful one increases buyer welfare by 43% in the costly treatment and 47% in the free treatment. We summarize our findings on buyer outcomes with our final result.

Result 5. Individual ratings can make buyers better or worse off, but on average ratings have no effect on buyer welfare.

5. Conclusion

We find evidence that non-contingent side-payments from sellers significantly influence rater behavior. Not only are ratings influenced by such payments, but in our data those payments are the sole determinant of what type of rating is sent. Despite existing evidence that altruism motivates raters to truthfully inform buyers (Lafky, 2014), in our environment we observe no such concern, with raters instead responding solely to the size of side-payments they receive. Raters are

willing to mislead buyers by recommending low-quality sellers who have made large transfers, and recommending against high-quality sellers who have made small transfers.

Our results suggest that those designing or participating in rating systems need to be mindful not only of explicit fraud, but also of biases introduced through preferential treatment of raters. For example, a consumer who is provided with a free review copy of a product may provide a more favorable evaluation of that product than if they were required to return the evaluation copy after they finished their review. While existing research on biased ratings has focused primarily on explicit forms of fraud in which sellers surreptitiously provide ratings for products (Dellarocas, 2006; Mayzlin et al., 2014), our results suggest explicit fraud is not the only source of intentionally skewed ratings. Implicit, reciprocal relationships unrelated to seller quality may also harm recommendations for future consumers.

Our study implements a simple, stylized experimental environment designed to cleanly isolate one component of the interaction between sellers and raters: the role of side-payments on choice of rating. Our results identify rating behavior when the rater is weighing reciprocity over side-payments versus altruism toward buyers. To isolate subject behavior, we omit some features found in rating environments in the field. First, our setting is a one-shot environment with one-sided ratings, which means that there is no opportunity for individual learning or concern for reputation building over time. Second, seller quality is determined exogenously, allowing us to focus solely on the role of side-payments as the source of reciprocity. Third, our raters do not "purchase" from the seller first-hand, meaning that they are not directly affected by seller quality. In the field, raters are frequently consumers who have already purchased from the seller, and are likely to be influenced by reciprocity over both the size of side-payments and the quality they personally experienced, in addition to potential altruism toward buyers. Finally, we note that the persuasiveness of side-payments in the field could vary considerably based on context: for example, raters may respond differently to side payments in anonymous online markets than in face-to-face interactions. Similarly, our design has a single buyer, however the number of consumers who observe ratings can vary widely by setting. Increasing the number of buyers might increase the altruistic value of a truthful rating and thereby reduce the relative influence of side-payments.

While our experiment was designed to explain the choices of raters, our results also provide some insights into seller behavior. We find no relationship between quality and the size of transfers when rating is free, but a small positive relationship between the two when rating is costly. The lack of relationship in the free treatment can be explained by sellers who believe that raters are concerned only with the size of transfer, and not with underlying quality. More difficult to explain is why we observe a positive relationship in the costly treatment. While the size of this effect is relatively modest, there is no obvious explanation for why higher quality sellers would give larger transfers, but only when ratings are costly. A future experimental design that identifies seller beliefs about raters, for example by informing sellers about raters' past histories of play, could help to explain the behavior we observed in our study.

Our parsimonious design is helpful for isolating the role of side-payments, however there are many opportunities for future studies to examine more complex settings. For example, a design that implemented endogenous choice of quality could build off of our results to provide insights into the dynamics of how ratings are used to reciprocate both quality and side-payments simultaneously. Even better would be to conduct similar experiments in the field, especially in a setting in which both seller quality and transfers between sellers and raters could be observed and easily quantified. Alternatively, it would be helpful to examine side-payments in settings with two-sided ratings, or when sellers and raters interact repeatedly, allowing for reputation-building and strategic considerations to develop over time.

Acknowledgments

We would like to acknowledge the helpful comments John Spraggon, Angela C. M. De Oliveira, Samuel Bowles, Abdul Kidwai, David Kingsley, attendees at the SEA, BEEMA, nd EEA conferences, and two anonymous referees during the writing and revision of the article. Any remaining errors are our own.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at 10.1016/j.socec.2019.101480 .

References

Abbink, K., Dasgupta, U., Gangadharan, L., Jain, T., 2014. Letting the briber go free: An experiment on mitigating harassment bribes. Journal of Public Economics 111 (C), 17–28.
Anderson, C.M., Putterman, L., 2006. Do non-strategic sanctions obey the law of demand? The demand for punishment in the voluntary contribution mechanism. Games and Economic Behavior 54 (1), 1–24. https://doi.org/10.1016/j.geb.2004.08.007.
Avery, C., Resnick, P., Zeckhauser, R., 1999. The market for evaluations. American Economic Review 564–584.
Balafoutas, L., Nikiforakis, N., Rockenbach, B., 2014. Direct and indirect punishment among strangers in the field. Proceedings of the National Academy of Sciences 111 (45), 15924–15927. https://doi.org/10.1073/pnas.1413170111.
Barr, A., Serra, D., 2010. Corruption and culture: An experimental analysis. Journal of Public Economics 94 (11–12), 862–869. https://doi.org/10.1016/j.jpubeco.2010.07.006.
Bolton, G., Greiner, B., Ockenfels, A., 2013. Engineering trust: Reciprocity in the production of reputation information. Management Science 59 (2), 265–285.
Burdìn, G., Halliday, S., Landini, F., 2018. The hidden benefits of abstaining from control. Journal of Economic Behavior & Organization. https://doi.org/10.1016/j.jebo.2017.12.018.
Carpenter, J.P., 2007. The demand for punishment. Journal of Economic Behavior & Organization 62, 522–542.
Charness, G., Rabin, M., 2002. Understanding social preferences with simple tests. The Quarterly Journal of Economics 117 (3), 817–869.
Chen, Y., Harper, M., Konstan, J., Li, S.X., 2010. Social comparisons and contributions to online communities: A field experiment on movielens. American Economic Review 100 (4), pp.1358–1398.
Dellarocas, C., 2006. Strategic manipulation of internet opinion forums: Implications for consumers and firms. Management science 52 (10), 1577–1593.
Dini, F., Spagnolo, G., 2009. Buying reputation on ebay: Do recent changes help? International Journal of Electronic Business 7 (6), 581–598.
Falk, A., Fischbacher, U., 2006. A theory of reciprocity. Games and Economic Behavior 54 (2), 293–315. https://doi.org/10.1016/j.geb.2005.03.001.
Falk, A., Kosfeld, M., 2006. The hidden costs of control. American Economic Review 96 (5), 1611–1630. https://doi.org/10.1257/aer.96.5.1611.
Fehr, E., Fischbacher, U., 2004. Third-party punishment and social norms. Evolution of Human Behavior 25, 63–87.
Fehr, E., Gächter, S., 2002. Altruistic punishment in humans. Nature 415 (10), 137–140.
Fischbacher, U., 2007. z-tree: Zurich toolbox for ready-made economic experiments. Experimental economics 10 (2), 171–178.
Fradkin, A., Grewal, E., & Holtz, D. (2019). Reciprocity in two-sided reputation systems: Evidence from an experiment on airbnb,.
Gneezy, U., 2005. Deception: The role of consequences. American Economic Review 95 (1), 384–394. https://doi.org/10.1257/0002828053828662.
Greiner, B., 2015. Subject pool recruitment procedures: Organizing experiments with orsee. Journal of the Economic Science Association 1 (1), 114–125. https://doi.org/10.1007/s40881-015-0004-4.
Kamei, K., Putterman, L., 2018. Reputation transmission without benefit to the reporter: A behavioral underpinning of markets in experimental focus. Economic Inquiry 56 (1), 158–172.
Klein, T., Lambertz, C., Spagnalo, G., Stahl, K., 2009. The actual structure of ebays feedback mechanism and early evidence on the effect of recent changes. International Journal of Electronic Business 7 (3), 301–320.
Klein, T.J., Lambertz, C., Spagnolo, G., Stahl, K.O., 2006. Last minute feedback. CEPR Discussion Paper No. 5693.
Lafky, J., 2014. Why do people rate? theory and evidence on online ratings. Games and Economic Behavior 87, 554–570. https://doi.org/10.1016/j.geb.2014.02.008.
Lafky, J., Wilson, A.J., 2015. Quality versus quantity in information transmission: Theory and experimental evidence. Working Paper. University of Pittsburgh, Department of Economics.
Li, L., Xiao, E., 2014. Money talks: Rebate mechanisms in reputation system design. Management Science 60 (8), 2054–2072.

Malmendier, U., Schmidt, K.M., 2017. You owe me. The American Economic Review 107 (2), 493–526.
Mayzlin, D., Dover, Y., Chevalier, J., 2014. Promotional reviews: An empirical investigation of online review manipulation. American Economic Review 104 (8), 2421–2455.
Nikiforakis, N., Mitchell, H., 2014. Mixing the carrots with the sticks: Third party punishment and reward. Experimental Economics 17 (1), 1–23. https://doi.org/10.1007/s10683-013-9354-z.
Nosenzo, D., Offerman, T., Sefton, M., van der Veen, A., 2015. Discretionary sanctions and rewards in the repeated inspection game. Management Science 62 (2), 502–517. https://doi.org/10.1287/mnsc.2014.2124.
Rabin, M., 1993. Incorporating fairness into game theory and economics. The American economic review 1281–1302.
Resnick, P., Zeckhauser, R., 2002. Trust among strangers in internet transactions: empirical analysis of ebay's reputation system. The economics of the internet and e-commerce. Emerald Group Publishing Limited, pp. 127–157.
Sutter, M., 2009. Deception through telling the truth?! experimental evidence from individuals and teams*. The Economic Journal 119 (534), 47–60. https://doi.org/10.1111/j.1468-0297.2008.02205.x.
Xiao, E., 2013. Profit-seeking punishment corrupts norm obedience. Games and Economic Behavior 77 (1), 321–344. https://doi.org/10.1016/j.geb.2012.10.010.
Ziegelmeyer, A., Schmelz, K., Ploner, M., 2012. Hidden costs of control: Four repetitions and an extension. Experimental Economics 15 (2), 323–340.