

## Manuscript Details

<b>Manuscript number</b>	JBEE_2019_216_R1
<b>Title</b>	Reciprocity through ratings: An experimental study of bias in evaluations
<b>Article type</b>	Research Paper

### 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.

<b>Keywords</b>	Ratings; Reciprocity; Punishment
<b>Taxonomy</b>	Behavioral Economics, Experiments in Economics
<b>Manuscript category</b>	Research Paper
<b>Corresponding Author</b>	Simon Halliday
<b>Corresponding Author's Institution</b>	Smith College
<b>Order of Authors</b>	Simon Halliday, Jonathan Lafky
<b>Suggested reviewers</b>	Jeffrey Carpenter, Kenju Kamei, Gary Bolton, Ben Greiner, Danila Serra

## Submission Files Included in this PDF

### File Name [File Type]

cover\_letter\_jbee\_halliday\_lafky\_revision.docx [Cover Letter]

JBEE revision response.pdf [Response to Reviewers]

Highlights.docx [Highlights]

manuscript\_revised.pdf [Manuscript File]

CombinedExpMat.pdf [Supporting File]

disclosure\_v2\_anon.docx [Supporting File]

JBEE\_appendix.pdf [Supporting File]

To view all the submission files, including those not included in the PDF, click on the manuscript title on your EVISE Homepage, then click 'Download zip file'.

## Research Data Related to this Submission

There are no linked research data sets for this submission. The following reason is given:  
We shall upload the data after a final editorial decision has been made.

18 August 2019

---

To Whom It May Concern,

Revise and Re-submit by Halliday and Lafky, REF: JBEE 2019 216

Please find in our submission the following materials:

- Revised manuscript without author details
- Appendix with additional theory and regressions based on the reviewer comments
- Reviewer responses document with point-by-point response to reviewers (including page references)
- Experiment instructions
- Disclosure Statement (anonymous)
- Data (to be submitted with Mendeley Data when the service is available)

*Online Appendix*

All data are currently provided and statistical analysis code will be made available with the accepted version of the paper to ensure reproducibility and transparency. These would be included in an online appendix at the time of publication. Informed Consent forms can also be provided with an accepted version of the paper, but are excluded for now as they include details that would reveal the authors' identities.

We also wanted to mention one additional piece of information. One of the authors (Halliday) is submitting his file for tenure in the fall, so any chance of a quick response from the reviewers would be greatly appreciated, but in no way expected.

Best regards,  
Simon Halliday  
Assistant Professor

# Response to Editor and Reviewers

Manuscript Reference: JBEE\_2019\_216

---

## Editor Comments

Both Reviewers liked the contribution of the paper and the design of the experiment, and found the results interesting. I concur with their assessment.

The Reviewers raise a series of concerns about the extent to which your experimental design is suitable to capture the setting of interest, and, more generally, about the generalizability of your findings.

I find these to be valid points that I encourage you to discuss in the paper, acknowledging, where necessary, the limitations of your experimental design.

The Reviewers also make a series of helpful suggestions about the data analysis and theoretical framework, which I would like you to consider carefully.

**Dear Editor and Reviewers,**

**Thank you for the insightful and constructive comments. We appreciate the opportunity to respond to the reviews and re-submit our manuscript. Your feedback was very valuable, and allowed us to significantly improve the paper. We have made substantial changes to all sections of the paper, but especially to the introduction and conclusion.**

**We address each of the reviewers' comments below. Reviewer comments are in plain text, while our responses are in boldface. We also include a new appendix that reports additional theoretical details and regression results.**

# Reviewer 1 Comments

1. The authors state that some examples of side-payments include free products (p. 3). This differs greatly from having pure monetary incentives (which is what is used in the experiment). While it is common to use monetary rewards as gifts in the experimental setup (e.g., in gift exchange experiments), I wonder to what extent the results in this particular setup can be generalized to non-monetary gifts. In practice, since sellers deliver products or services that are evaluated by the raters, the “gifts” can potentially be perceived as part of the products and services provided by the seller. Hence, these gifts may provide an additional signal of the seller’s quality. For example, a seller providing gifts may be perceived to be of higher quality because these gifts are viewed as part of the services or products they deliver to their buyers.

**Thank you for pointing this out, it is an important distinction to make clear about our environment. Our intention is to refer to sellers who are not just generous with their customers, but who specifically favor reviewers with free products. We have rewritten footnote 3 (on p. 3) to clarify the types of interaction between firm and reviewer we are discussing. In particular, we emphasize that we are referring to reviewers who “have a more favorable experience than can reasonably be expected for the general customer.”**

2. The model makes the assumption (p. 5, footnote 5) that buyers ignore raters’ reciprocity concerns towards the sellers. That is, if a buyer observes a good rating, the authors assume that this implies the seller has a high quality and therefore the buyer will follow the rater’s recommendation. This assumption is very simplistic as it ignores strategic concerns. If a buyer is aware that sellers can provide side payments to raters, and that raters have reciprocity concerns towards sellers, then logically the buyer should take this into account when making decisions as to whether to follow the rater’s

recommendation. However, the authors do not account for this in their model. I question to what extent the theoretical predictions rely on this assumption.

**Thank you for the suggestion, this is a point we should have explained more carefully. We have considerably lengthened footnote 6 (on p. 6) to explain why our assumption of credulousness is relatively minor, and what would happen if buyers were fully strategic instead. In short, credulousness is reasonable even for a fully strategic buyer, so long as they believe ratings to be even slightly better than noise. If they do not believe ratings to be informative, then they should ignore the ratings entirely, which results in a setting where no ratings are sent and no information is transferred. “Babbling” outcomes are always a possibility in these types of communication environments, even in a setting with no possibility of side-payments.**

3. I think an interesting result which the authors have not focused on is the fact that sellers do not seem to act strategically. I would have expected low-quality sellers to make larger transfers (in fact, this is also what the model predicts), but the authors do not appear to find this result. Table 2 in fact shows that this effect is acting in the opposite direction (the estimated coefficient on “Quality” is positive). This shows that something else is influencing the sellers’ transfer decisions other than what is predicted in the model. The authors should provide an explanation for this.

**We have added more discussion of seller behavior in the results section (p. 10 and p. 13), and in the conclusion (p. 20). We have also added further regression results in Table A2 of the appendix. In short, we find no relationship between quality and transfer in the free treatment, but a positive relationship between quality and transfer in the costly treatment. The free treatment behavior can be explained by sellers who believe raters are unconcerned with quality, however we do not have a simple explanation for seller behavior in the costly treatment. We suggest in the conclusion (p. 21) that this may be a good topic for future research.**

4. In the conclusion, the authors mention that the “desire to reciprocate side-payments effectively eliminates any concern raters have for informing buyers of a seller’s true quality” (p. 18). However, the current paper does not provide any evidence of any baseline concern that raters have towards buyers. If this comparison is made in reference to results from other papers (e.g., Lafky 2014), the authors should make this clear.

**We agree that this comparison should have been clearer. We now reference Lafky (2014) in the introduction (p. 3) and conclusion (p. 19) to make clear that there is existing evidence of rater altruism toward buyers.**

5. Similarly, the point on the nature of interactions affecting the impact of side-payments is not the focus of the study. The paper studies a one-shot interaction where each seller is matched with only one rater and one buyer. I am not sure the authors can use this specific setup and results to make any strong conjectures on the impact of side payments for small versus large businesses (p. 19).

**Thank you for suggesting this. After re-reading that paragraph we agree that our initial interpretation was too broad. In rewriting the conclusion we have struck all reference to large and small businesses, and now make a softer claim that the side-payments may have different effects in different settings (on p. 20):**

**“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.”**

6. A related work on reviews and reciprocity that the authors should look at is “Reciprocity in Two-sided Reputation Systems: Evidence from an Experiment on Airbnb” (by Andrey Fradkin, Elena Grewal, and David Holtz)

**Thank you, we have included this paper in the introduction in our discussion of the role of reciprocity (toward the end of p. 2).**

7. The authors make a comparison to Engel (2011)'s survey article on p. 12. Given that many factors within the dictator game environment itself can influence dictator giving, I am not sure whether one can say that "transfers from sellers do appear to be motivated by pure altruism" just by comparing average giving in their setup and the average as given in the survey. A more valid comparison would have been one where the authors conduct a separate within-subject dictator game with similar parameters as their main task (which they do not do in their experiment).

**After re-reading that section, we agree that the comparison was tenuous. We have cut that paragraph entirely, and instead transition directly to discussion of Table 2.**

#### **Reviewer 1 minor comments:**

1. The comparative static result of the model seems selective. The authors only consider the two cases of  $r = G$  vs.  $r = \emptyset$  and  $r = B$  vs.  $r = \emptyset$ . What about  $r = G$  vs.  $r = B$ ?

**Thank you for catching this. Footnote 5 now explains that, in order to give a particular rating, the rater needs to prefer giving that rating to *both* no rating and the opposite rating. As we now explain in the theory appendix, comparing only  $r = G$  vs.  $r = \emptyset$  (and likewise only  $r = B$  vs.  $r = \emptyset$ ) is a sufficient condition, as the condition to switch to  $r = \emptyset$  will always bind before the condition for switching to the other rating (i.e.,  $r = G$  vs.  $r = B$ ).**

**We are also now more explicit in characterizing rater behavior, now making clear on p. 7 when a rater will prefer to give  $r = \emptyset$  rather than either  $r = G$  or  $r = B$ .**

2. The statement “This difference should not have affected subject behavior” in footnote 7 is not true. In the model, the subjects’ behavior is contingent on their expectation of the quality draw (e.g.,  $E[q]$ ), which depends on the underlying support.

**Thank you for catching this, and apologies for our incorrect statement. Our original intention was to imply that, because this is a one-shot game there was no opportunity for subjects to learn the empirical distribution over time, but our claim was absolutely false. We now simply state that “The effect on subject behavior should have been minimal.” (footnote 8, p. 8).**

3. The summary statistics in the bottom half of Table 1 is not discussed anywhere in the paper.

**Thank you for catching this omission. We have reformatted Table 1, in accordance with Reviewer 1’s comment 4 below. We now briefly discuss both earnings and demographics on p. 11.**

4. In general, the summary statistics are very difficult to follow. The authors cite averages throughout the main text (e.g., % of sellers who transfer positive amounts on p. 10, the average % transfers by sellers on p. 12, average transfers preceding positive and negative ratings on p. 13, and the proportion of misleading ratings on p. 16, etc.). It would be very useful to have these presented in a table separate by treatments for the reader to follow.

**We have edited Table 1 on p. 11 to report a wide array of statistics on subject behavior. We now break the summary statistics up by role within the experiment, which we believe makes the table easier to understand.**



5. Figures 1 and 2 are also quite difficult to read. I think it makes more sense to have the data for both treatments on the same graph, and then have separate panels/graphs for “unrated”, “positive”, and “negative” ratings.

**Thank you for the suggestion. We have included a new three-panel Figure 1 to replace the previous Figures 1 and 2 (on p. 12). We have also adjusted the formatting of the figure slightly in order to make it easier to read.**

6. In Table 2, the authors do not explain or discuss the fact that the statistical significance of “Quality” in columns (1) and (2) once they include interaction terms.

**Thank you for pointing out this omission. We now explain Table 2 in more detail on p. 10, as discussed above in our response to Reviewer 1 comment 3. In short, we see a statistically significant relationship between quality and transfer in the costly treatment, but no such relationship in the free treatment. We report regressions restricted by treatment in the appendix in Table A1 (Columns 1 and 2 for the Free treatment and columns 3 and 4 for the Costly treatment).**

7. On p. 13, the authors should consider making “Results are very similar following an ordered probit... increase in the likelihood that the rating is favorable” a footnote discussion. It took a few re-reads before it became obvious that these are robustness checks, and that the results are not included anywhere in the main paper or appendices.

**Thank you for the suggestion. We now include ordered probit regression results in the appendix in Tables A3 and A4, and have added the following to page 14:**

**“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.”**

8. Table 3: What is the interpretation of the (marginally statistically significant) coefficients for |Quality – 8| and |Transfer – 8|?

**Thank you for catching this omission. We have added the following while discussing Table 3:**

**“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.” (p. 14)**

9. The authors should provide  $R^2$  for any OLS regression outputs.

**Our apologies for omitting the  $R^2$  in the original paper. We now report the Adjusted  $R^2$  as a measure of goodness of fit for each of the OLS regressions in the main paper and in the appendix.**

10. The paragraph on the top of p.17: the authors state that “We report the absolute frequency of misleading ratings...”. However, no statistics or figures were given anywhere in the paper, and the paragraph itself did not state any frequencies.
- Thank you for the suggestion. We originally intended this paragraph to reference the analysis for Result 3, however we realize now that we did a**

**poor job of making that reference clear. We have moved the paragraph the reviewer quoted to two paragraphs below Result 3, and rewritten the first sentence to begin “Although we observe that approximately one-third of ratings are misleading...” (p. 17) We believe that these edits make clearer our original intended meaning.**

## **Reviewer 2 Comments**

Concerns regarding the contribution of this paper to the existing literature.

1. Some papers dealing with rating systems in the context of online commerce are missing and should be added. In particular several studies have underlined the limitations of rating systems (Dellarocas et al., 2006, Dini and Spagnolo, 2009). Dellarocas et al. (2006) and Klein et al. (2005, 2007) provided empirical evidence of reprisals and reciprocity at work on eBay, which ultimately serves to artificially increase the number of positive ratings and reduce the number of negative ratings (see also Resnick and Zeckhauser, 2002).

**Thank you for all of the helpful citations. We have added references to these papers at various points in the introduction and conclusion, including p. 2 (Resnick and Zeckhauser), p. 4 (Dini and Spagnolo 2009), p. 2 (Klein, 2006, 2009), p. 3 and p. 19 (Dellarocas, 2006). We must add one caveat, though: as far as we could tell, Klein (2006) is an updated version of Klein (2005) and Klein (2009) is the published version of the Klein (2007) working paper. If other references were intended, we will gladly review their inclusion.**

2. I would suggest the authors to show more explicitly how their paper contributes to the existing literature. from my point of view they should highlight two main findings :

- a. first, the fact that quality and transfers are not substitute but rather complement in the seller's decision to transfer a positive amount of money to the rater.

**Thank you for the suggestion. As mentioned above in the response to Reviewer 1, Comment 3, we now discuss this behavior in more detail in both the Results and Conclusion sections of the paper, as well as including regression results in a new appendix (see Table A1). We clarify that seller behavior differs by treatment, showing no relationship between quality and transfer in the free treatment, but a positive relationship (i.e., complements) in the costly treatment.**

- b. Second, the fact that the seller quality does not influence the rater's decision, indicating that rater's intrinsic motivation to help consumers is relatively low.

**Thank you for highlighting this contribution; we agree that it is one of the main results in the paper. Consequently, we are now more explicit about the fact that our results suggest raters exhibit little concern for buyers in our setting. For example, in the conclusion (p. 18) we now state “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.”**

3. **Concerns regarding the experimental design.** Some design choices are not obvious and should be discussed in more detail. First, two key components of

most (online) rating systems are missing in the current paper. In this current paper the rater differs from the seller or the buyer. This sharply contrast with the functioning of most rating systems where the consumers or the sellers themselves are the raters. For instance in C2C (customer-to-customer) online marketplaces like eBay or AmazonMarketplace, the huge number of participants makes it difficult to centralize monitoring. This is the reason why online market places let traders monitor and rate each other's transactions by providing them with a decentralized reputation-building mechanism. The eBay Feedback Forum that is claimed to be a crucial component of the success of eBay, is a good example of such a decentralized mechanism. The idea is that traders themselves are often in a better position to monitor and evaluate their partners than the marketplace. In the eBay forum, both buyers and sellers have the opportunity to rate each other. The buyer can send a "positive", "negative", or "neutral" rating and the seller a "positive" rating. I would recommend the authors to discuss this point in the discussion section.

**Thank you for pointing this out. We have substantially rewritten the introduction and conclusion to better explain the limitations of our experimental design. We now emphasize that our rater does not “buy” from the seller on p. 4 and p. 19. We are also explicit that we implement a one-sided rating system, in which only sellers receive ratings. We make clear in footnote 7 on p. 8 in the section on the experimental design that our design eliminates the possibility of reputational concerns entirely. We also separate out the role of reputation in two-sided markets vs. the markets that we investigate on p. 2-3. We furthermore clarify the implications of results and the abstraction away from markets with reputation on p. 19.**

4. A second key component of online systems is that such rating systems are based on reputation. This dimension is missing here. Indeed the game is one shot. At least the authors should justify more this design choice.

**Please see our response to comment 3, immediately above. We agree that reputation is a part of many online interactions, as documented in the citations above, however our design is intended to abstract from the strategic opportunities that reputational concerns can create. We have attempted to make this distinction clearer in the introduction and conclusion.**

5. **Concerns regarding the theoretical framework** I would suggest the author to present the straightforward theoretical standard predictions in absence of any social preferences before presenting predictions with other regarding preferences. By working backward, the buyer should be indifferent between the two options X and Y. Since rating is costly the rater should never rate. Anticipating this, the seller should never sent any amount to the rater. From my point of view the authors should attempt to identify the rater's reciprocity concerns toward the seller coupled with intrinsic motivation to help the buyer in her choice. Regarding reciprocity, I have some concerns regarding how the author attempts to incorporate reciprocity in the rater's utility function. Indeed I'm not totally convinced by the way reciprocity is incorporated in the utility function. Why not having considered Rabin's model of reciprocity?

**Thank you for these suggestions. We have considerably rewritten the theory section, in response to this comment and that of Reviewer 1. We now characterize behavior with standard self-interested preferences prior to introducing social preferences for our full analysis.**

**We have also eliminated the functional form for our reciprocity term, leaving it simply as a generic function of the seller's transfer. Our only**

requirement is that the reciprocity term leads to positive reciprocity for sufficiently large transfers, and negative reciprocity for sufficiently small transfers. We explain this in footnote 4 and mention there are various models of reciprocity, including Rabin (1993), that are compatible with our theory.

6. Concerns regarding the data analysis.

- a. I would suggest the authors to estimate the probability that the buyer will choose the seller.

Thank you for the suggestion. We now report regression results estimating the probability that the buyer will choose the seller in Table A2 of the appendix and comment on the results in the main manuscript in Footnote 10 on p. 18 in reference to Result 4. In the regressions reported in Table A2, we use two different specifications. In columns 1 (OLS) and 2 (Probit), we estimate the regressions with demographic characteristics and whether the buyer receives any rating (good or bad). We see that simply receiving any rating does not result in a statistically significant increase in the probability that the buyer chooses the seller. However, in columns 3 (OLS) and 4 (Probit) we estimate similar regressions but instead include a dummy variable for receiving a *positive* rating, and we see that the coefficient is large, positive, and statistically significant meaning that a positive rating results in a greater likelihood that the buyer will choose the seller relative to a negative rating.

- b. Table 2. Why not controlling for risk aversion in the estimates.

**Thank you for suggesting this, but unfortunately we are not able to control for risk aversion as we did not collect information on risk preferences in the post-experimental survey. It would be interesting to know how much seller risk preferences influenced their willingness to “gamble” with larger transfers to raters, but when we designed the experiment we wanted to investigate mainly the rater and buyer behavior rather than seller behavior. Because of that design choice, seller risk preferences were not a necessary component of our original research question and design.**

- c. Please report the number of censored observations in the tobit estimates;

**Thank you for catching this omission. We have added the number of censored observations to the notes in Table 2 in the paper (p. 13), and Table A1 in the appendix.**

- d. An interesting finding is that the quality variable captures a positive and significant coefficient in table 2 indicating that transfers and quality are not substitute but rather complement. I would suggest the authors to discuss more this finding. Why is it the case? Is it due to the experimental design? Would it be different if the buyer could receive a proxy about the seller quality?

**Thank you for suggesting this. Please see our responses to Reviewer 2, Comment 2a and Reviewer 1, Comment 3.**



We study the effect of side payments on ratings

We find a large influence of side payments, which fully crowd out concern for consumers

Firms and regulators should be mindful of implicit, as well as explicit relationships between sellers and raters.

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

Simon D. Halliday\*      Jonathan Lafky<sup>†</sup>

## 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. (JEL D47, D82, D83)

Keywords: Ratings, Reciprocity, Punishment

This work was supported by research funds from Smith College, Lafayette College and Carleton College.

---

\*Smith College Department of Economics, Pierce Hall 107, 21 West St, Northampton, MA, 01062, shallday@smith.edu, Phone: (413) 585-3529

<sup>†</sup>Corresponding Author. Carleton College Department of Economics, Northfield, MN 55057. jlafky@carleton.edu. Phone: (507) 222-4103

# I 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.<sup>1</sup> 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, and 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 et al. 2010; Li and 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 and 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 and Gächter 2002; Fehr and Fischbacher 2004; Nikiforakis and Mitchell 2014; Balafoutas, Nikiforakis, and Rockenbach 2014; Nosenzo et al. 2015).<sup>2</sup> Reciprocity has been demonstrated to play a role in rating behavior in the field, in settings such as Airbnb (Fradkin, Grewal, and Holtz 2019) and eBay (Resnick and 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

---

<sup>1</sup>U.S. Census Bureau, U.S. Retail Trade Sales - Total and E-commerce (1998-2016). Retrieved from <https://www.census.gov/data/tables/2016/econ/arts/annual-report.html>

<sup>2</sup>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 and Putterman 2006), third-party punishment games (Carpenter 2007) and in reviewing contexts (Lafky 2014).

other members of the market and to protect one’s own reputation (Klein et al. 2006; Klein et al. 2009; Bolton, Greiner, and 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 et al. 2014), or petty bribes between officials and citizens (Barr and 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 and 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.<sup>3</sup> 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 reci-

---

<sup>3</sup>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.

procuity 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 and 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, and Chevalier 2014), or by exploiting a good reputation to sell low-quality products (Dini and 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.

## II 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  $r = \emptyset$ .

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  $\pi$  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

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

where  $\alpha(a)$  and  $\beta$  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.<sup>4</sup> The rater therefore exhibits either positive or negative concern for

---

<sup>4</sup>Different models of reciprocity compatible with our environment include Rabin (1993), Falk and Fischbacher (2006), and Charness and Rabin (2002). Our concern is only with the qualitative predictions of the reciprocity

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,  $\beta$ , 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:<sup>5</sup>

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

For simplicity, we assume that the buyer treats ratings credulously and randomizes over the choices if no rating is sent.<sup>6</sup> 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, Equation 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:

---

function, that raters exhibit positive reciprocity for sufficiently high transfers, and negative reciprocity for sufficiently small transfers.

<sup>5</sup>In order to give a positive rating, the rater actually needs to prefer  $r = G$  to both  $r = \emptyset$  and  $r = B$ , however the condition for preferring  $r = G$  to  $r = \emptyset$  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  $r = \emptyset$ . These comparisons are made explicit in the appendix.

<sup>6</sup>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.

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

or equivalently:

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

The intuition conveyed by Equation 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 Equations 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.*

Equations 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 Equation 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.

### III 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.<sup>7</sup> Participants were randomly assigned to groups of three 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.<sup>8</sup> 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.

---

<sup>7</sup>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 Burdin, Halliday, 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.

<sup>8</sup>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.

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:

$$\begin{aligned} \text{Seller} \quad \pi_S &= \begin{cases} \$32 - a & \text{if chosen by buyer} \\ \$16 - a & \text{if not chosen by buyer} \end{cases} \\ \text{Rater} \quad \pi_R &= \begin{cases} a - c & \text{if rater chose to send a message} \\ a & \text{if rater chose not to send a message} \end{cases} \\ \text{Buyer} \quad \pi_B &= \begin{cases} X & \text{if buyer chose seller} \\ Y & \text{if buyer chose outside option} \end{cases} \end{aligned}$$

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.

## IV 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.

Figure 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

TABLE 1: Summary statistics.

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

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

Standard deviations in parentheses.

FIGURE 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.

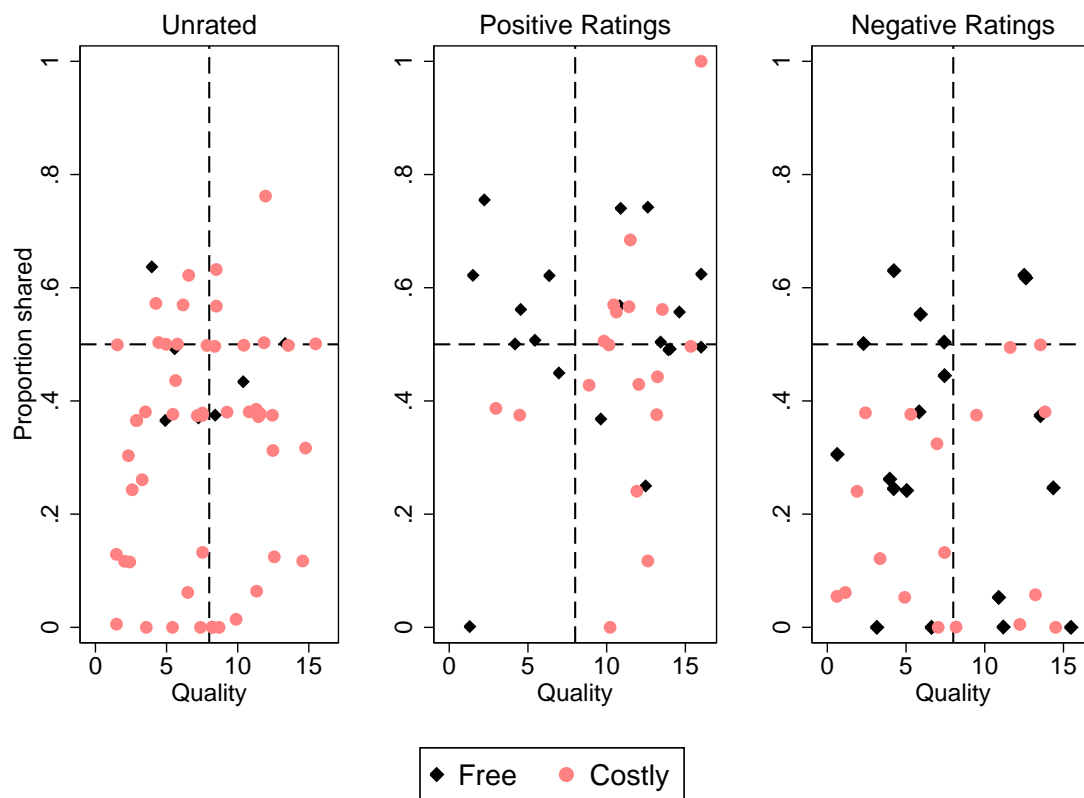


TABLE 2: Amount transferred by seller

	(1) Tobit	(2) OLS	(3) Tobit	(4) OLS
Cost	-1.646** (0.688)	-1.579** (0.607)	-2.886* (1.516)	-2.773** (1.316)
Quality	0.167** (0.0794)	0.156** (0.0701)	0.0752 (0.132)	0.0672 (0.115)
Female	-0.889 (0.680)	-0.901 (0.595)	-0.835 (0.679)	-0.850 (0.596)
Age	-0.169 (0.189)	-0.142 (0.164)	-0.170 (0.184)	-0.144 (0.160)
Quality $\times$ Cost			0.148 (0.163)	0.143 (0.144)
Constant		8.740** (3.392)		9.516*** (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.

\*  $p < 0.10$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$ .

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 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.<sup>9</sup> 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

---

<sup>9</sup>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.

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*** (0.0666)	-0.426*** (0.0782)	-0.589*** (0.172)	-0.625*** (0.168)	-0.646*** (0.177)	-0.643*** (0.156)
Quality	0.0144* (0.00875)	0.0142 (0.00884)	0.00382 (0.0129)	0.00203 (0.00905)	0.000568 (0.0135)	-0.00184 (0.00819)
Transfer	-0.00367 (0.0105)	-0.00308 (0.0108)	-0.0121 (0.0148)	-0.00784 (0.0100)	0.00458 (0.0199)	0.00958 (0.0128)
Female	-0.123 (0.0756)	-0.131 (0.0807)	-0.126* (0.0756)	-0.133 (0.0816)	-0.123* (0.0742)	-0.138* (0.0819)
Age	0.00374 (0.0241)	0.00367 (0.0243)	0.00306 (0.0239)	0.00405 (0.0238)	0.00892 (0.0246)	0.00799 (0.0262)
Quality $\times$ Cost			0.0143 (0.0171)	0.0197 (0.0161)	0.0156 (0.0172)	0.0219 (0.0158)
Transfer $\times$ Cost			0.00978 (0.0197)	0.00531 (0.0183)	0.0203 (0.0225)	0.00865 (0.0168)
Quality $- 8$					0.0356** (0.0171)	0.0351* (0.0183)
Transfer $- 8$					0.0400 (0.0256)	0.0326* (0.0185)
Constant		0.721 (0.510)		0.848* (0.490)		0.457 (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 (0.101)	0.0806 (0.104)	-0.157 (0.311)	-0.0308 (0.277)
Quality	0.0151 (0.0105)	0.0160 (0.0113)	0.00837 (0.0135)	0.0115 (0.0155)
Transfer	0.0630*** (0.0109)	0.0668*** (0.0135)	0.0554*** (0.0191)	0.0646*** (0.0208)
Female	-0.0272 (0.0982)	-0.0469 (0.101)	-0.0308 (0.0979)	-0.0485 (0.104)
Age	0.0235 (0.0381)	0.0256 (0.0389)	0.0265 (0.0380)	0.0253 (0.0398)
Quality $\times$ Cost			0.0165 (0.0227)	0.0108 (0.0238)
Transfer $\times$ Cost			0.0146 (0.0333)	0.00203 (0.0291)
Constant		-0.587 (0.847)		-0.527 (0.927)
$N$	74	74	74	74
Adjusted $R^2$		0.236		0.216

Robust standard errors in parentheses. Probit coefficients are marginal effects.

\*  $p < 0.10$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$ .

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 Figure 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 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 difference 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.<sup>10</sup> 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:

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

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

---

<sup>10</sup>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 5: Buyer Outcomes

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

Standard errors in parentheses.

\*  $p < 0.10$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$ , via Mann-Whitney.

$\pi_{B,F} = 8.78$  in the free treatment, meaning that in both treatments the ex-ante buyer payoffs are very close to the a seller’s ex-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.*

## V 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, Dover, and Chevalier 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.

## References

- Abbink, Klaus et al., “Letting the briber go free: An experiment on mitigating harassment bribes”, *Journal of Public Economics* 111(C), 2014, 17–28.
- Anderson, Christopher M. and Louis Putterman, “Do non-strategic sanctions obey the law of demand? The demand for punishment in the voluntary contribution mechanism”, *Games and Economic Behavior* 54(1), 2006, 1–24.
- Avery, Christopher, Paul Resnick, and Richard Zeckhauser, “The market for evaluations”, *American Economic Review*, 1999, 564–584.
- Balafoutas, Loukas, Nikos Nikiforakis, and Bettina Rockenbach, “Direct and indirect punishment among strangers in the field”, *Proceedings of the National Academy of Sciences* 111(45), 2014, 15924–15927.
- Barr, Abigail and Danila Serra, “Corruption and culture: An experimental analysis”, *Journal of Public Economics* 94(11–12), 2010, 862–869.
- Bolton, Gary, Ben Greiner, and Axel Ockenfels, “Engineering trust: reciprocity in the production of reputation information”, *Management Science* 59(2), 2013, 265–285.
- Burdin, Gabriel, Simon Halliday, and Fabio Landini, “The hidden benefits of abstaining from control”, *Journal of Economic Behavior & Organization*, 2018.
- Carpenter, Jeffrey P., “The demand for punishment”, *Journal of Economic Behavior & Organization* 62, 2007, 522–542.
- Charness, Gary and Matthew Rabin, “Understanding social preferences with simple tests”, *The Quarterly Journal of Economics* 117(3), 2002, 817–869.
- Chen, Yan et al., “Social Comparisons and Contributions to Online Communities: A Field Experiment on MovieLens”, *American Economic Review* 100(4), 2010, pp. 1358–1398.
- Dellarocas, Chrysanthos, “Strategic manipulation of internet opinion forums: Implications for consumers and firms”, *Management science* 52(10), 2006, 1577–1593.
- Dini, Federico and Giancarlo Spagnolo, “Buying reputation on eBay: Do recent changes help?”, *International Journal of Electronic Business* 7(6), 2009, 581–598.
- Falk, Armin and Urs Fischbacher, “A theory of reciprocity”, *Games and Economic Behavior* 54(2), 2006, 293–315.
- Falk, Armin and Michael Kosfeld, “The Hidden Costs of Control”, *American Economic Review* 96(5), 2006, 1611–1630.
- Fehr, Ernst and Urs Fischbacher, “Third-party Punishment and Social Norms”, *Evolution of Human Behavior* 25, 2004, 63–87.
- Fehr, Ernst and Simon Gächter, “Altruistic Punishment in Humans”, *Nature* 415(10), 2002, 137–140.

- Fischbacher, Urs, “z-Tree: Zurich toolbox for ready-made economic experiments”, *Experimental economics* 10(2), 2007, 171–178.
- Fradkin, Andrey, Elena Grewal, and David Holtz, “Reciprocity in Two-sided Reputation Systems: Evidence from an Experiment on Airbnb”, 2019.
- Gneezy, Uri, “Deception: The Role of Consequences”, *American Economic Review* 95(1), 2005, 384–394.
- Greiner, Ben, “Subject pool recruitment procedures: organizing experiments with ORSEE”, *Journal of the Economic Science Association* 1(1), 2015, 114–125.
- Kamei, Kenju and Louis Putterman, “Reputation transmission without benefit to the reporter: A behavioral underpinning of markets in experimental focus”, *Economic Inquiry* 56(1), 2018, 158–172.
- Klein, TJ et al., “The actual structure of eBay’s feedback mechanism and early evidence on the effect of recent changes”, *International Journal of Electronic Business* 7(3), 2009, 301–320.
- Klein, Tobias J et al., “Last minute feedback”, *CEPR Discussion Paper No. 5693*, 2006.
- Lafky, Jonathan, “Why do people rate? Theory and evidence on online ratings”, *Games and Economic Behavior* 87, 2014, 554–570.
- Lafky, Jonathan and Alistair J Wilson, *Quality versus quantity in information transmission: Theory and experimental evidence*, Working Paper, University of Pittsburgh, Department of Economics, 2015.
- Li, Lingfang and Erte Xiao, “Money Talks: Rebate Mechanisms in Reputation System Design”, *Management Science* 60(8), 2014, 2054–2072.
- Malmendier, Ulrike and Klaus M Schmidt, “You owe me”, *The American Economic Review* 107(2), 2017, 493–526.
- Mayzlin, Dina, Yaniv Dover, and Judith Chevalier, “Promotional Reviews: An Empirical Investigation of Online Review Manipulation”, *American Economic Review* 104(8), 2014, 2421–2455.
- Nikiforakis, Nikos and Helen Mitchell, “Mixing the carrots with the sticks: third party punishment and reward”, *Experimental Economics* 17(1), 2014, 1–23.
- Nosenzo, Daniele et al., “Discretionary Sanctions and Rewards in the Repeated Inspection Game”, *Management Science* 62(2), 2015, 502–517.
- Rabin, Matthew, “Incorporating fairness into game theory and economics”, *The American economic review*, 1993, 1281–1302.



- Resnick, Paul and Richard Zeckhauser, “Trust among strangers in Internet transactions: Empirical analysis of eBay’s reputation system”, in: *The Economics of the Internet and E-commerce*, Emerald Group Publishing Limited, 2002, 127–157.
- Sutter, Matthias, “Deception Through Telling the Truth?! Experimental Evidence From Individuals and Teams\*”, *The Economic Journal* 119(534), 2009, 47–60.
- Xiao, Erte, “Profit-seeking punishment corrupts norm obedience”, *Games and Economic Behavior* 77(1), 2013, 321 –344.
- Ziegelmeyer, Anthony, Katrin Schmelz, and Matteo Ploner, “Hidden costs of control: four repetitions and an extension”, *Experimental Economics* 15(2), 2012, 323–340.

*[For review: Costly treatment]*

**Instructions**

**Introduction**

Welcome to this experiment in the economics of decision-making. Please read these instructions carefully as they explain how you earn money from the decisions you make in this experiment. There will be no talking during today's session. If you have a question, please raise your hand and an experimenter will answer your question in private.

All participants in today's session will be randomly and anonymously assigned to groups of three participants. The participants in your group will be known only as Participant A, Participant B and Participant C. You are equally likely to be assigned each of these three roles. Other than these labels, you will not learn the identity of any member of your group during the experiment or after it is completed.

At the end of today's session you will be paid in cash and in private based on your decisions, and the decisions of the other participants in your group.

**Participant A**

- The experiment will begin with the computer randomly picking a number, called the X number, between 0 and 16 for Participant A.
- All whole numbers between 0 and 16 are equally likely to be picked.
- Participant A will be assigned \$16.00 and must decide how much of the \$16.00 to give to Participant B in increments of \$1.00.

**Participant B**

- Participant B begins by learning the X number that was picked for Participant A, and learning the amount of money that was given to Participant B by Participant A.
- After learning both of these numbers, Participant B is given the option to pay \$0.50 to send a message to Participant C.
- Participant B can choose to send the message "Choose Participant A," the message "Do not choose Participant A" or to not send any message.

**Participant C**

- Participant C begins by learning what message, if any, was sent by Participant B.
- Next, Participant C is given the option to either choose Participant A, or choose another randomly drawn number, called the Y number.

- As was the case with the X number, the Y number is drawn between 0 and 16 with all whole numbers in that range being equally likely to be drawn.

### **Earnings**

- If Participant C chooses Participant A, then Participant C receives \$1.00 times the X number, while Participant A receives \$16.00.
- If participant C chooses the Y number, then participant C receives \$1.00 times the Y number, and participant A receives no additional payment.

For each participant, earnings amount to the following:

#### *Participant A*

- + \$16.00 at the start
- Amount of money assigned to Participant B by Participant A
- + \$16.00 if Participant C chooses Participant A

#### *Participant B*

- + Amount of money assigned to Participant B by Participant A
- \$0.50 if Participant B chooses to send a message

#### *Participant C*

- + Participant A's X number times \$1.00 ( $X \times \$1.00$ ) if Participant C chooses Participant A
- + A randomly assigned Y number times \$1.00 ( $Y \times \$1.00$ ) if Participant C does not choose Participant A

### **Payment**

At the end of the experiment you will be paid in cash and in private based on the amount of money you have earned. You will receive a \$5.00 show-up fee plus whatever money you earned during the experiment. **[If you lost money during the experiment, your losses will be subtracted from your show-up fee.]**

### **Comprehension Quiz**

We will now take a few minutes for a brief quiz. This quiz is intended solely to check that you have followed these instructions and understand the rules of the experiment. Your answers to these questions will have no effect on your earnings. The numbers used in these questions are

purely illustrative and should not be taken as suggestive of behavior you will experience in the experiment.

**Your Computer Number:** \_\_\_\_\_

**Question 1.**

What is the largest possible X number? What is the smallest possible X number?

**Question 2.**

What does Participant B learn about Participant A prior to choosing a message?

**Question 3.**

What are the possible messages that can be sent by Participant B?

**Question 4.**

If Participant C chooses Participant A, how does the X number affect the amount of money they each earn?

*[For review: Free treatment]*

## **Instructions**

### **Introduction**

Welcome to this experiment in the economics of decision-making. Please read these instructions carefully as they explain how you earn money from the decisions you make in this experiment. There will be no talking during today's session. If you have a question, please raise your hand and an experimenter will answer your question in private.

All participants in today's session will be randomly and anonymously assigned to groups of three participants. The participants in your group will be known only as Participant A, Participant B and Participant C. You are equally likely to be assigned each of these three roles. Other than these labels, you will not learn the identity of any member of your group during the experiment or after it is completed.

At the end of today's session you will be paid in cash and in private based on your decisions, and the decisions of the other participants in your group.

### **Participant A**

- The experiment will begin with the computer randomly picking a number, called the X number, between 0 and 16 for Participant A.
- All whole numbers between 0 and 16 are equally likely to be picked.
- Participant A will be assigned \$16.00 and must decide how much of the \$16.00 to give to Participant B in increments of \$1.00.

### **Participant B**

- Participant B begins by learning the X number that was picked for Participant A, and learning the amount of money that was given to Participant B by Participant A.
- After learning both of these numbers, Participant B is given the option to send a message to Participant C.
- Participant B can choose to send the message "Choose Participant A," the message "Do not choose Participant A" or to not send any message.

### **Participant C**

- Participant C begins by learning what message, if any, was sent by Participant B.
- Next, Participant C is given the option to either choose Participant A, or choose another randomly drawn number, called the Y number.

- As was the case with the X number, the Y number is drawn between 0 and 16 with all whole numbers in that range being equally likely to be drawn.

## **Earnings**

If Participant C chooses Participant A, then Participant C receives \$1.00 times the X number, while Participant A receives \$16.00. If participant C chooses the Y number, then participant C receives \$1.00 times the Y number, and participant A receives no additional payment.

For each participant, earnings amount to the following:

### *Participant A*

- + \$16.00 at the start
- Amount of money assigned to Participant B by Participant A
- + \$16.00 if Participant C chooses Participant A

### *Participant B*

- + Amount of money assigned to Participant B by Participant A

### *Participant C*

- + Participant A's X number times \$1.00 ( $X \times \$1.00$ ) if Participant C chooses Participant A
- + A randomly assigned Y number times \$1.00 ( $Y \times \$1.00$ ) if Participant C does not choose Participant A

## **Payment**

At the end of the experiment you will be paid in cash and in private based on the amount of money you have earned. You will receive a \$5.00 show-up fee plus whatever money you earned during the experiment.

### **Comprehension Quiz**

We will now take a few minutes for a brief quiz. This quiz is intended solely to check that you have followed these instructions and understand the rules of the experiment. Your answers to these questions will have no effect on your earnings. The numbers used in these questions are purely illustrative and should not be taken as suggestive of behavior you will experience in the experiment.

**Your Computer Number:** \_\_\_\_\_

#### **Question 1.**

What is the largest possible X number? What is the smallest possible X number?

#### **Question 2.**

What does Participant B learn about Participant A prior to choosing a message?

#### **Question 3.**

What are the possible messages that can be sent by Participant B?

#### **Question 4.**

If Participant C chooses Participant A, how does the X number affect the amount of money they each earn?

## Comprehension questions

Please solve the following comprehension questions.

They have no consequence on your earnings and only serve to check if all participants in the experiment have understood the rules.

Question 1: Participant A is randomly assigned an  $X$  number of 5. Participant A chooses to allocate \$8.00 to Participant B. Participant B does not send a message to Participant C. Participant C chooses Participant A.

Earnings of Participant A:	<input type="text" value="8"/>
Earnings of Participant B:	<input type="text" value="0"/>
Earnings of Participant C:	<input type="text" value="0"/>

Question 2: Participant A is randomly assigned an  $X$  number of 5. Participant A chooses to allocate \$5.00 to Participant B. Participant B does not send a message to Participant C. Participant C does not choose Participant A and is randomly assigned a  $Y$  number of 8. What are the incomes?

Earnings of participant A:	<input type="text" value="5"/>
Earnings of participant B:	<input type="text" value="0"/>
Earnings of participant C:	<input type="text" value="0"/>

Question 3: Participant A is randomly assigned an  $X$  number of 4. Participant A chooses to allocate \$12.00 to Participant B. Participant B sends a message to Participant C. Participant C chooses Participant A. What are the incomes?

Earnings of participant A:	<input type="text" value="12"/>
Earnings of participant B:	<input type="text" value="0"/>
Earnings of participant C:	<input type="text" value="0"/>

OK



Your X value is:

1.00

How much of your \$16.00 would you like to transfer to Participant B?

Please enter the amount you wish to transfer:

Submit

Participant A's X number is:

1.00

The amount of money sent to you by Participant A is:

5.00

Please choose which message, if any, you would like to send to Participant C.

Your choice:

- ☐ Send the message: "Choose Participant A"
- ☐ Send the message: "Do not choose Participant A"
- ☐ Do not send a message

Submit

Participant B sent you the message: "Do not choose Participant A."

Please decide if you will choose Participant A.

If you choose Participant A you will receive Participant A's X number.

If you do not choose Participant A you will receive the Y number.

Your choice: ☐ Choose Participant A  
☐ Do not choose Participant A

Submit

## Demographics

What is your gender?

- ☐ male  
☐ female

What is your age?

When did you start studying at the university (mm/yyyy), e.g. 09/2013?

What subject(s) are you majoring in or do you plan to major in?

Approximately how many experiments have you previously participated in?

[continue](#)

To Whom It May Concern,

Disclosure Statement for Submission to JBEE

*Financial support for the experiments*

- The experiments were conducted with supporting funds from the authors' employers.

*Conflict of Interest*

- The authors declare that they have no relevant or material financial interests that relate to the research described in this paper.
- No external body served as a reviewer for this paper prior to submission or publication.
- The authors have no relevant paid or unpaid external positions.

*IRB Clearance*

- IRB clearance was obtained and the IRB confirmation number can be provided on request.

Best regards,

The authors (anonymous as this statement is provided to reviewers)

# Appendix

Reciprocity through ratings:

An experimental study of bias in evaluations

## A Theory Appendix

To characterize rater behavior we need only compare utilities when  $r = G$  versus  $r = \emptyset$  and when  $r = B$  versus  $r = \emptyset$ . To see this, observe first that the requirement for a rater to give a positive rating rather than a negative rating is

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

Or equivalently

$$\alpha(a) \cdot \pi + \beta \cdot (q - E(k)) > 0 \tag{1}$$

Compare this to the condition for giving a positive rating rather than no rating:

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

The sum of the rater's concern for seller and concern for the buyer must simply be positive in the first case, but must larger than  $2c$  in the second case. In other words, equation 2 implies equation 1. This means that, as the rater's utility (the sum of the  $\alpha$  and  $\beta$  terms) falls, the condition to switch to  $r = \emptyset$  will always bind before the condition to switch to  $r = B$ . Thus to understand when the rater will give  $r = G$  we need only consider  $r = G$  versus  $r = \emptyset$ .

A nearly identical argument shows that, as the sum of the  $\alpha$  and  $\beta$  terms rises, the condition for the rater to switch from  $r = B$  to  $r = \emptyset$  will always bind before the condition for switching from  $r = B$  to  $r = G$ .

## B Additional Regressions

TABLE A1: Amount transferred by seller, restricted by treatment.

	(1)	(2)	(3)	(4)
	Tobit Free	OLS Free	Tobit Costly	OLS Costly
Quality	0.130 (0.118)	0.117 (0.106)	0.220** (0.100)	0.206** (0.0886)
Female	-3.308*** (1.065)	-3.023*** (0.948)	0.438 (0.836)	0.233 (0.729)
Age	-0.198 (0.314)	-0.187 (0.284)	-0.0841 (0.245)	-0.0616 (0.209)
Constant		10.88* (5.447)		4.632 (4.339)
$N$	45	45	87	87
Adjusted $R^2$		0.160		0.0303

Standard errors in parentheses

Censoring at 0: 5 (Free) and 13 (Costly), Censoring at 16:  $n = 1$  (Costly).

\*  $p < 0.10$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$

TABLE A2: Probability that buyer follows rater recommendations.

	(1)	(2)	(3)	(4)
	OLS	Probit	OLS	Probit
Cost	-0.0457 (0.104)	-0.0442 (0.101)	-0.0549 (0.0600)	-0.0528 (0.0615)
Female	0.142 (0.0891)	0.141* (0.0848)	0.0258 (0.0610)	-0.00212 (0.0493)
Age	0.0177 (0.0136)	0.0205 (0.0186)	0.0316 (0.0204)	0.0280* (0.0161)
Received Any Rating ( $r \neq \emptyset$ )	-0.0343 (0.0987)	-0.0347 (0.0961)		
Positive Rating ( $r = G r \neq \emptyset$ )			0.861*** (0.0566)	0.444*** (0.0692)
Constant	0.101 (0.305)		-0.590 (0.427)	
$N$	132	132	74	74
Adjusted $R^2$	-0.000589		0.702	

Standard errors in parentheses

\*  $p < 0.10$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$



TABLE A3: Rating type estimated by ordered probit without interaction terms.

	(1)	(2)	(3)	(4)
	Coefficients	Do not Choose	No Rating	Choose
Cost	0.241 (0.248)	-0.0708 (0.0738)	-0.000892 (0.00713)	0.0717 (0.0721)
Quality	0.0419 (0.0261)	-0.0123* (0.00747)	-0.000155 (0.00130)	0.0125* (0.00759)
Transfer	0.139*** (0.0325)	-0.0410*** (0.00837)	-0.000516 (0.00429)	0.0415*** (0.00857)
Female	0.00507 (0.200)	-0.00149 (0.0588)	-0.0000188 (0.000712)	0.00151 (0.0595)
Age	0.00815 (0.0651)	-0.00240 (0.0192)	-0.0000302 (0.000348)	0.00243 (0.0194)
Cutoff 1	0.807 (1.441)			
Cutoff 2	2.131 (1.450)			
$N$	132	132	132	132

Standard errors in parentheses

\*  $p < 0.10$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$

TABLE A4: Rating type estimated by ordered probit with interaction terms.

	(1) Coefficients	(2) Do not Choose	(3) No Rating	(4) Choose
Cost	1.047 (0.990)	-0.302 (0.272)	-0.00956 (0.0317)	0.311 (0.281)
Quality	0.0451 (0.0559)	-0.0130 (0.0158)	-0.000412 (0.00138)	0.0134 (0.0162)
Transfer	0.229** (0.0935)	-0.0660*** (0.0232)	-0.00209 (0.00683)	0.0681*** (0.0249)
Female	0.0293 (0.200)	-0.00845 (0.0577)	-0.000268 (0.00175)	0.00872 (0.0593)
Age	0.0215 (0.0644)	-0.00620 (0.0185)	-0.000197 (0.000874)	0.00640 (0.0191)
Quality×Cost	0.000296 (0.0596)	-0.0000852 (0.0172)	-0.00000270 (0.000546)	0.0000879 (0.0177)
Amount Received×Cost	-0.127 (0.0973)	0.0365 (0.0261)	0.00116 (0.00391)	-0.0376 (0.0275)
Cutoff 1	1.733 (1.884)			
Cutoff 2	3.062 (1.925)			
<i>N</i>	132	132	132	132

Standard errors in parentheses

\*  $p < 0.10$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$