Why Do People Rate? Theory and Evidence on Online Ratings.

Jonathan Lafky*

Abstract

The rapid growth of online retail in the last decade has led to widespread use of consumer-generated ratings. This paper theoretically and experimentally identifies influences that drive consumers to rate products and examines how those factors can create distortions in product ratings. By manipulating payoffs and effectively "deactivating" either the buyer or seller side of an artificial laboratory market, raters' behavior is decomposed into buyer-centric and seller-centric components. The cost of providing a rating also plays a major role in influencing rating behavior, with high and low quality sellers being rated more often than those of moderate quality.

Keywords: Online Ratings, Altruism, Punishment.
JEL classification: C91, D64, D83, L86.

*Lafayette College, Easton, PA 18042. lafkyj@lafayette.edu. Phone: (610) 330-5314 Fax: (610) 330-5715. This paper was previously titled "Why Do People Write Reviews? Theory and Evidence on Online Ratings."

1 Introduction

Internet commerce is a large and rapidly growing component of the economy. Internet retail accounted for \$224.3 billion in sales for 2012, up 16.2% from 2011. Typical growth over the past decade has been even higher, averaging approximately 20% annually. Online retail's share of total U.S. retail has also increased tremendously over the past decade, climbing from just over 1% in 2001 to 5.2% by the end of 2012.[1]

The rapid growth and popularity of internet retail is not surprising. Virtually any good can be purchased on the internet, in every model, style, or color produced. The enormous selection offered to consumers means that they must often choose between several goods with similar observable characteristics but potentially different levels of quality. Without firsthand experience, it may be difficult or impossible for consumers to tell which of several similar-looking products is of the highest quality.

In an effort to alleviate this problem and to encourage sales, many internet retailers provide customer-based rating and review systems for their products. In these systems, consumers (sometimes restricted only to previous buyers) are allowed to post written reviews as well as numerical scores for products. These ratings are then made available to future buyers to inform them of the product's qualities, allowing them to make more informed purchases. For example, Amazon.com allows customers to leave ratings between one and five stars in one-star increments, as well as written comments about products they have purchased.

The average review score can vary considerably for products that have otherwise similar characteristics, and may be the only insight consumers have into a product's unobservable qualities before they buy. As Chevalier and Mayzlin (2006) demonstrate, ratings can significantly influence buyers' behavior and have a substantial impact on the success or failure of a product. But why are ratings given in the first place? Are people taking time to give these ratings in order to help their anonymous fellow shoppers, or are they writing out of gratitude or anger that they feel towards online merchants? Are raters equally likely to evaluate all products, or do they speak up only if they have a strong opinion? This paper examines possible motivations for the provision of numerical ratings in a theoretical framework and then isolates those motivations in an experimental setting.

To preview the results, I find evidence that consumers are motivated by concern for both buyers and sellers when they decide to rate products. Making rating less attractive through the introduction of a small cost has a large effect on the volume and distribution of ratings.

[1] U.S. Census Bureau December, 2013 Monthly Retail Trade Report. Retrieved from http://www.census.gov/retail/

Ratings in the presence of a cost take on a U-shaped distribution, which can lead to average ratings that are not representative of true quality. A possible solution to this problem is to provide small discounts to consumers who provide ratings, thereby compensating for any inconveniences or opportunity costs associated with rating products.

This paper focuses solely on numerical ratings and does not examine textual reviews. While consumers may ultimately be influenced by written reviews, they are not incorporated in the numerical score that is typically the first signal consumers receive as to a product's quality. Additionally, for the same reason that written reviews may be useful, they are also difficult to analyze in a rigorous and objective manner. Just as written reviews express nuances not easily captured through a numerical score, the reviews themselves are not easily quantified without imposing substantial subjectivity. Nonetheless, some of this paper's insights on consumer rating behavior may be applicable to written reviews as well. In particular it seems plausible that polarization may occur in written reviews for the same reasons that it occurs in numerical ratings. Extending these results to the domain of written reviews would be a useful area for future research.

The remainder of the paper is organized as follows. Section **??** surveys relevant past research. Section **??** provides motivating data and poses the basic questions to be addressed. Section **??** introduces a theoretical framework for analyzing rating behavior and isolating concern for sellers from concern for buyers. Section **??** lays out the experimental design and hypotheses. Section **??** presents results from the experiments while section **??** discusses implications of those findings. Section **??** concludes.

2 Related Literature

There is a small but growing literature on online ratings, with most existing work focusing on how consumers are influenced by ratings and how well those ratings can predict market outcomes. Much of the literature takes the existence of ratings as a given, avoiding the questions of why or how accurately the ratings are created. Chevalier and Mayzlin (2003), for example, convincingly show that ratings influence consumers' book purchasing behavior, but they do not examine whether the ratings accurately reflect book quality.

Other authors have questioned the value of ratings, arguing that they may predict future sales without actually influencing them. Duan et al. (2005) and Dellarocas et al. (2004) argue that consumer-generated ratings for movies are simply a gauge of underlying word-of-mouth communication, rather than a driver of movie success or failure. Such arguments over causality illustrate one of the major advantages of moving ratings research into the laboratory, where it is easier to cleanly identify causal relationships between rater incentives, ratings and purchases. Bolton et al. (2004) use such an experimental setting to show a causal relationship between feedback and trustworthiness of sellers in a simulated market. In their setting, however, feedback is automatic, and thus is not subject to any potential biases or strategic behavior that raters may exhibit in practice.

There is a small body of existing research on behaviors that may influence ratings. This work largely focuses on the potential for biases from self-selection, as in Li and Hitt (2007) and Hu et al. (2009). Self-selection across time is explored by Li and Hitt, who consider possible distortions in ratings for newly released products on Amazon.com. They find that products experience consistent rising and falling patterns across time, which can be explained by early adoption among "avid fans" and a later backlash from average consumers. Hu et al. uses similar insights to explain the tendency for long-term ratings to take a J-shaped distribution. They propose two biases to explain the distribution: A "moan and groan" underreporting bias, and a self-selection "purchasing" bias, similar to that proposed by Li and Hitt. Li and Xiao (2010) also find a bias in rating behavior, with consumers being more sensitive to the cost of rating for high quality sellers than for low quality sellers.

Self-selection is not the only issue relevant to rating generation, however, and Wang (2010) demonstrates that other factors such as social identity and anonymity can play major roles in consumers' decision to provide ratings. He finds that a strong sense of social identity considerably increases the quantity and quality of ratings. In a similar vein, Chen et al. (2008) use social comparisons to encourage users of MovieLens, a movie recommendation website, to rate more movies. They show that providing users with information on how their rating output compares to others' substantially increases the volume of ratings. They also find some evidence that a user's propensity for altruism predicts their likelihood of rating movies that have few existing ratings. This is significant for the current paper, as it suggests that at least some users are motivated by altruism when providing ratings.

It is important to distinguish the current line of research from several papers that have been written on two-sided reputation systems. Houser and Wooders (2005), for example, examine the impact of reputations in eBay auctions, in which buyers and sellers rate one another. As evidenced by eBay's change in 2008 to a one-sided rating system, in which buyers may rate sellers but sellers cannot rate buyers, two-sided systems can introduce the undesirable possibility of strategic rating behavior. In contrast, consumers in the one-sided system considered in this paper need not worry about being punished or rewarded for their ratings. They can rate products based solely on their own opinions.

3 Motivating Data

As a motivating example of online ratings, data was collected from the Amazon.com website in November 2008. The distributions of ratings for more than 400 products were collected, encompassing more than 17,500 separate ratings. While most previous research in this area has focused on books, music and movies, all of which have relatively subjective quality, the current data was drawn from the "Home Improvement" section of the website, which includes products such as lawn mowers, flashlights and electric chainsaws. These products were chosen under the assumption that, as tools intended to solve specific problems, they would exhibit more objective quality than creative works such as books.[2]

Figure **??** shows the distribution of individual ratings for products with different average ratings. For reference, customers at Amazon.com can give ratings from one star to five stars, in one-star increments. Only one item (less than 0.25% of all products) had an average rating of less than two stars, and thus is not included.

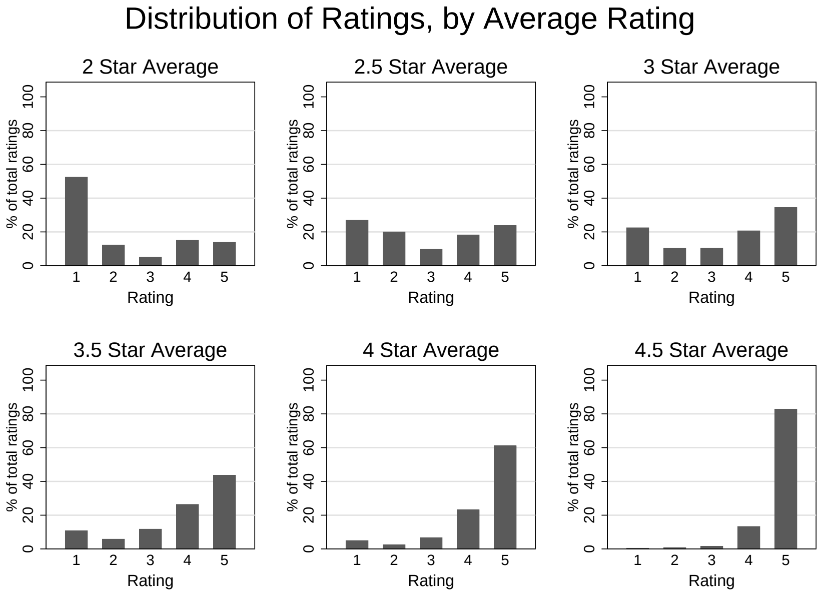

Figure 1: Ratings distributions by average rating.

Looking at the distributions, one feature is particularly striking: middling ratings, especially ratings of two stars and three stars, are uncommon even among products with *average* ratings of two or three stars. The reasons behind these distributions are not clear, however,

[2] The effects described below are found for other product categories, including those with more subjective quality. Anecdotally, products with more subjective quality, such as books and movies, seem to suffer from more polarization than those of objective quality. This is an interesting area for future research.

as several distinct mechanisms could generate the same pattern. The simplest explanation would be that quality itself tends towards extremes, where products realize binary qualities of "success" and "failure" with little else in between. It is also possible, however, that the pattern comes not from the underlying distribution of quality, but from the motivations that drive consumers to provide ratings.

If people view the act of rating a product to be intrinsically burdensome but they nonetheless want to help other buyers, the same pattern could emerge. Acceptable but unremarkable products would not be rated because the benefit to the rater from informing others would be smaller than the cost of providing the rating. High or low quality products would be rated because the rater could have a large impact on other buyers' welfare.[3]

Alternatively, raters may take the time to rate in an attempt to punish or reward sellers for their quality, as in Levine (1998). A buyer who receives a defective product may seek retribution against the good's seller by damaging their reputation with a negative rating. Likewise, a buyer who is pleased with a recently purchased good may give the seller a positive rating as a reward or encouragement for their high quality. In both cases the reaction elicited from the buyer is intense enough to outweigh any costs of rating. Products of moderate quality, however, would elicit neither reward nor punishment.

While this paper works to distinguish between a rater's intrinsic concerns for buyers and sellers, other motivations certainly generate some of the ratings we see online, such as consumers enjoying the very act of rating. It is also likely that some consumers behave strategically, giving a rating to encourage sellers to provide them with high quality products in the future. While these and other motivations are surely the source of some ratings in the field, our goal is to isolate concern that raters have for each side of the market, not to fully explain the complicated interactions between raters, buyers and sellers.

Despite identifying each of these possible reasons for rating, it remains unclear what is actually driving consumers to rate. Due to the difficulty of distinguishing between the motivations using only the field data, we move into the lab to obtain control that can be difficult to achieve in a field setting.

4 Theory

This section develops a theoretical model to formalize the insights described above. It characterizes behavior for buyers and sellers interacting in a simple, stylized market and generates predictions that can be tested in the laboratory.

[3] This corresponds to the reporting bias discussed in Hu et al. (2009).

4.1 The Model

In this model there are n buyers, $B_1, B_2 \dots B_n$, each deciding from which of two sellers, S_1 and S_2 to buy.[4] Buyer B_1 will be referred to as the "first buyer" and $B_2, \dots, B_n$ will be known as "second buyers." At the start of the game, each seller chooses a quality level, $q_i \in [0, q_{max}]$, for some $q_{max} \in \mathbb{R}^+$.[5] The choice of q_i is a seller's private information. After sellers choose their qualities, buyer B_1 selects one of the sellers. Since B_1 has no information to distinguish one seller from the other, for ease of notation assume that B_1 chooses S_1. B_1 learns S_1's quality, q_1 and is then given the opportunity to pay a cost of $c \in \mathbb{R}_0^+$ in order to provide a rating $r \in [\underline{r}, \overline{r}] \subset \mathbb{R}$ for S_1.[6] This cost can be interpreted as both the opportunity cost of rating as well as any effort a consumer expends from the act of rating.

After B_1 makes his rating decision, all other buyers learn what rating, if any, B_1 gave. If B_1 did not give a rating, the other buyers cannot tell which seller B_1 selected. Finally, $B_2, \dots, B_n$ each simultaneously select one of the sellers.

Sellers' payoffs are $U_{S_i}(q_i, n_i) = n_i \cdot (\overline{u} - aq_i)$ where $\overline{u} \in \mathbb{R}^+$ is the utility from providing the minimum quality, $a \in \mathbb{R}^+$ is the marginal cost of quality, and $n_i \in \mathbb{N}_0^+$ is the number of buyers who selected S_i. B_1's payoffs are given by $U_{B_1}(q) = bq - I_r c$, where $b \in \mathbb{R}^+$ is the marginal benefit of quality and I_r is an indicator function for whether B_1 rated or not. The payoffs for all other buyers are simply $U_{B_i}(q) = bq$.

Given this framework the unique equilibrium is for sellers to set the minimum quality of $q = 0$, and for B_1 to never provide a rating so long as $c > 0$. This model does not reflect observed behavior, however, in that there are tens of millions of buyer-generated ratings on the internet. In order to explain this discrepancy I extend the model in the spirit of Fehr and Gächter (2000) to include regard for others.[7] The timing of the game remains the same, however B_1's payoffs are rewritten as

$$U_{B_1}(q_1) = bq_1 + \alpha \cdot (q_1 - R)U_{S_1}(q_1, n_1) + \beta \sum_{i=2}^{n} U_{B_i}(q_{j(i)}) - I_r c \tag{1}$$

[4] Two is the minimum number of sellers that prevents unrealistic signalling behavior. With a single seller, not rating has the potential to convey as much information as rating.

[5] For simplicity prices are normalized to zero. This is done to remove the possibility that prices would be used as a signal for quality, which would complicate the task of inferring rater's motivations. One interpretation of this model is an analysis of products at a given price, meaning that quality can be thought of as value for money.

[6] Although we permit a broad range of possible ratings, only two ratings are necessary from a theoretical perspective, representing "buy" and "don't buy" messages. The richer message space is included here, and in the experimental design below, for consistency with commonly used online rating systems.

[7] An alternate explanation for voluntary rating, especially in the realm of non-durable goods, is a repeat-customer motive in which a consumer rates to improve their own future interactions with a merchant. Such a motivation may drive some online ratings, although it is outside of the scope of this paper.

$\alpha \in \mathbb{R}$ and $\beta \in \mathbb{R}$ measure B_1's concern for sellers and buyers, respectively; $R \in \mathbb{R}^+$ is a quality level that B_1 considers to be fair treatment; and $q_{j(i)}$ is the quality corresponding to whichever seller j is selected by buyer i. To understand what it means for B_1 to judge R to be "fair," consider his reactions to deviations of q_1 away from R. If $q_1 < R$, S_1 was selfish in choosing quality, meaning that $q_1 - R < 0$ and thus B_1 has vengeful concern for his seller. When $q_1 > R$, S_1 was generous in choosing quality, and B_1 has altruistic concern for his seller. When $q_1 = R$, then $q_1 - R = 0$ and B_1 is unconcerned by S_1's utility. Note that B_1's altruism or hostility towards his seller becomes more intense the further S_1's choice becomes from B_1's opinion of fair quality. It should also be noted that "fair" quality in this case is defined as the quality that elicits neither altruism nor hostility from B_1. Fairness might reasonably instead be defined as behavior that would elicit a positive, altruistic response from a buyer, but it is not modeled as such in this paper.

Sellers' preferences remain the same, and for simplicity do not include other-regarding components. Introducing other-regarding preferences for sellers would alter the quality levels provided, but would not qualitatively affect the analysis of first or second buyer behavior. Buyer behavior is characterized below for the full range of possible qualities, and thus all variations in quality levels are already accounted for.

The addition of other-regarding preferences to the model gives a reasonable starting point for describing behavior, although it is not ideal for a meaningful analysis. As was the case with the motivating data, concern for sellers and second buyers still cannot be separately identified from the first buyer's actions. Isolating these motivations requires adding one additional feature to the model.

Prior to the beginning of the game, nature randomly determines if it will be a *buyer-fixed* or *seller-fixed* game. The type of game is known only to B_1, although sellers and second buyers know that it will be either buyer-fixed or seller-fixed with equal probability. A buyer-fixed game has the same structure as the previously described model, except that $B_2, \ldots, B_n$ receive fixed payoffs of $f \in \mathbb{R}$, independent of the actions taken by any player. Similarly, in a seller-fixed game, all buyers receive their normal payoffs while the sellers receive payoffs of f, independent of any player's action. In this way, it is possible to "deactivate" either sellers or second buyers from B_1's decision to rate, as B_1 is affected only by his concern for sellers or second buyers in each role's respective game type.[8] This means that in a buyer-fixed game B_1's rating decision is determined entirely by his value of α and the quality he receives. Likewise in a seller-fixed game, B_1's rating decision is affected only by his value of β and his received quality. Note that, because they cannot tell which type of game is being played, all

[8] For certain environments an alternate approach can be found in Servátka (2009), in which motivations for altruistic behavior are isolated by manipulating the information available to subjects.

players other than B_1 behave the same in both types of games.

One setting that is not examined is a "none-fixed" environment, in which neither the buyers nor sellers receive fixed payments. While this setting is certainly the most natural, it introduces complications that cloud the view of a first buyer's motivation for rating. In such a setting, a second-buyer cannot be certain that the first buyer will rate with second buyers' welfare in mind. For example, if all sellers offer relatively low quality, but the seller chosen by B_1 is the highest of these low quality sellers, second buyers would benefit from the first buyer giving a positive rating. The first buyer, however, may instead give a negative rating, as he wishes to harm the seller. As a result of these conflicting motivations, ratings become less informative for second buyers, who become less inclined to follow them. This communication breakdown in turn leads to a change in the benefit of giving a rating for a first buyer. These interactions, while interesting, are removed from the goal of separately identifying concern for buyers from concern for sellers, and are thus not considered in this paper.

4.2 First Buyer Behavior

The first buyer's behavior can be characterized through two propositions. The details of the propositions, in particular the derivations of $\underline{q}_S$ and $\overline{q}_S$, are explained in the mathematical appendix.

Proposition 1. *In the buyer-fixed game, the first buyer will rate a seller if and only if* $q \in [0, \underline{q}_S) \cup [\overline{q}_S, q_{max}]$*, for some cutoff values* $\underline{q}_S$*,* $\overline{q}_S$*.*

Intuitively, Proposition 1 says that, in the presence of positive rating costs, the first buyer will only rate sellers of sufficiently high or low quality. The cutoffs $\underline{q}_S$ and $\overline{q}_S$ depend upon R, α and c, and show that the first buyer will only rate sellers with quality far enough above or below the "fair" level, R. Figure **??** shows the cutoff values varying with α, with all quality levels between $\underline{q}_s$ and $\overline{q}_s$ going unrated.

The range of unrated qualities is decreasing in the level of concern raters have for other buyers, α, and increasing in the cost of rating, c. This is a key insight in explaining the U-shaped distribution of ratings. If there is no cost of rating ($c = 0$), all quality levels will be rated, but with any positive rating cost ($c > 0$), a "blind spot" of unrated qualities emerges centered around R.

Behavior in the seller-fixed game is similar, with one important difference. Because each second buyer can potentially select any of the sellers, B_1's utility in a seller-fixed round can be influenced not only by S_1's quality, but also by the quality of the seller who has not yet been selected. B_1 thus needs to have beliefs about the quality that buyers will receive if

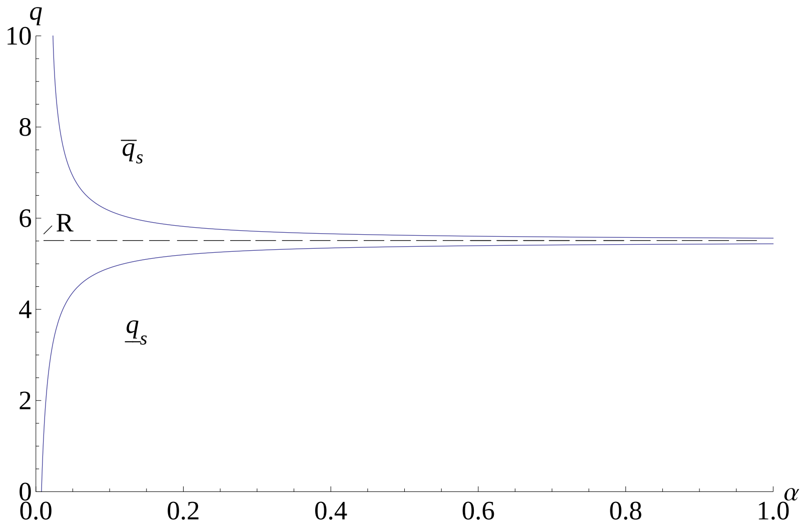

Figure 2: Cutoffs for buyer-fixed rating, with $n = 3, c = .25, a = .36, \overline{u} = 6$ and $R = 5.5$

they switch from S_1 to S_2. Denote the quality that B_1 believes to be offered by S_2 by q'. Note that no assumptions are made about the source of q', allowing for the possibility that it corresponds to the actual quality of the other seller but not requiring it to do so. The cutoff values $\underline{q}_B$ and $\overline{q}_B$ below are functions of β and q', as shown in the mathematical appendix.

Proposition 2. *In the seller-fixed game, the first buyer will rate a seller if and only if* $q \in [0, \underline{q}_B) \cup [\overline{q}_B, q_{max}]$ *for some cutoff values* $\underline{q}_B$, $\overline{q}_B$.

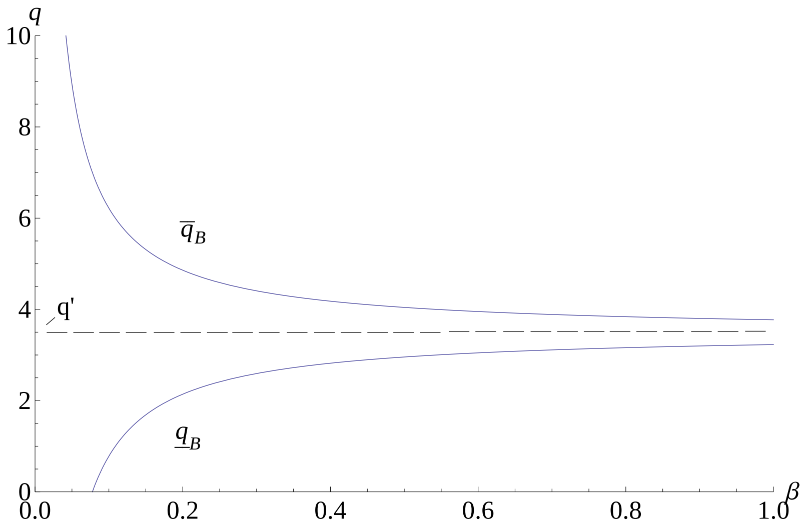

Figure 3: Cutoffs for seller-fixed rating, with $n = 3$, $c = .25$, $b = .92$ and $q' = 3.5$.

Figure **??** shows the cutoff values for different values of β. Similar to the setting for buyer fixed rounds, all qualities between $\underline{q}_B$ and $\overline{q}_B$ will not be rated, and this gap is decreasing in β and increasing in c.

The intuition here is similar to Proposition 1, except that the range of unrated qualities is now centered around q' instead of R. In other words, when a rater is concerned with helping other buyers, they only rate sellers who are sufficiently far above or below average quality. As in the buyer-fixed game the range of unrated qualities is increasing in the cost of rating and decreasing in the level of concern for other buyers.

5 Experimental Design

Despite having theoretical predictions for rating behavior, it is difficult to test these predications against online ratings in the field. In analyzing data from ratings websites, it is difficult to control for product quality or cost of rating, two variables essential to identifying behavior. These problems can be overcome by moving to the laboratory, where it is possible to perfectly control both quality and the cost of rating.

This paper uses a novel experimental design intended to isolate subjects' motivations for giving ratings. The experiment was conducted using Fischbacher's (2007) z-Tree software over networked computers in the Pittsburgh Experimental Economics Laboratory. A total of 240 subjects were recruited from the student populations of the University of Pittsburgh and Carnegie Mellon University. Each session consisted of 20 subjects with no prior knowledge of the experiment. Each session began with the distribution of written instructions that were then read aloud to all subjects. A brief comprehension quiz was administered, subjects played 20 rounds of the experiment and then completed a brief questionnaire. All experimental materials are included at the end of the paper. The instructions used in the Costly and Free treatments were nearly identical, though the minimal differences between the two are indicated at the beginning of the experimental appendix.

This experiment utilizes a 2 x 2 design, with the treatment variables being the cost of rating and the payment structure within each round. While the cost of rating varied between subjects, with each subject seeing only a single treatment, the type of round varied within subjects, with a single subject being exposed to both treatments.

At the end of each session, subjects were paid in cash for one randomly selected round. Sessions lasted one hour or less and average earnings were approximately \$11.00, including a \$5.00 show-up fee. At the beginning of each of the 20 rounds, subjects were randomly and anonymously assigned into four groups, each consisting of five players. Within each group subjects were randomly assigned roles, with two subjects taking the role of sellers, one subject in the role of first buyer and two subjects in the role of second buyers. These roles were also randomly assigned at the beginning of each of the 20 rounds. Due to the combination of random assignment and anonymity, subjects could not tell if they had interacted previously with their groupmates in any given round.

The sellers in each group were the first to act, each choosing an integer quality level from 0 to 10, inclusive. Next, and without knowing the sellers' qualities, the first buyer chose one of the sellers to "purchase" from. The first buyer then learned the quality of his chosen seller and was given the option to provide that seller with a rating. He was only able to rate the seller whose quality he had observed, and was not allowed to rate the unchosen seller. The cost of giving a rating varied by treatment; it was either \$0.25 (Cost treatment) or \$0.00 (Free treatment). A rating consisted of an integer score from 1 to 5, inclusive. To ensure that first buyers in the costly treatment would never lose money, each subject was also given a \$1.00 "round completion fee" at the end of each round.

After the first buyer decided whether to give a rating, the second buyers were informed of what, if any, rating was given. If no rating was given, the second buyers could not tell which seller the first buyer picked. After seeing what, if any, rating was given the second buyers each simultaneously selected a seller for themselves. Sellers with quality level q received payoffs of $\$6.10 - \$0.34q$ each time a buyer picked them. First and second buyers who chose a seller with quality q received payoffs of $\$0.92q$.

Each round was selected with equal probability by the computer to be either seller-fixed or (second) buyer-fixed. In a seller-fixed round, all sellers received \$6.00, regardless of what decisions were made. Likewise, in a buyer-fixed round, all second buyers received \$6.00 total, independent of any subjects' decisions. All subjects knew that each round would be either seller-fixed or buyer-fixed, but only the first buyer knew the round's type while he made his decisions. Sellers and second buyers learned the round type only at the end of the round, after their decisions had been made. Sellers and second buyers were faced with the same decision and incentives in each type of round, even though their actions would only affect their payoffs 50% of the time.

By implementing this payoff and information structure, this design effectively "deactivates" either sellers or second buyers as targets for the first buyer's concern. For example, in a seller-fixed round the first buyer cannot influence his seller's payoffs in any way, because the seller will only receive the fixed payment of \$6.00. To ensure that first buyers had no influence over sellers, ratings were visible to sellers in buyer-fixed rounds, but not in seller-fixed rounds. This was done to exclude the possibility that first buyers would rate negatively in a seller-fixed round in order to express their displeasure directly to sellers, as demonstrated in Xiao and Houser (2008).[9]

[9] Xiao and Houser show that responders in an ultimatum game accept lower offers when they are provided

The payment structure (seller-fixed versus buyer-fixed) is varied within subject to ensure that sellers and second buyers behave the same in both round types. While the within-subject design may have greater potential for an experimenter demand effect by making subjects more aware of which side of the market would be affected by a rating, it does not suggest anything about how each side should be treated. An experimenter demand effect might increase the overall frequency of rating, but it should not bias the relative frequency of each type of round.

A between-subject design would not be practical, as subjects who received the fixed payments (e.g., sellers in a buyer-fixed session) would learn that their actions did not affect their payoffs. This in turn could be resolved by keeping subjects in the same roles throughout the experiment and not informing sellers or second buyers of which type was fixed until the end of the session. This, however, would introduce additional self-interested incentives for first buyers to rate, as they could punish low quality sellers in early rounds in order to receive higher quality from those same sellers in later rounds. While such a self-interested incentive is certain to drive some ratings in the field, its presence clouds our ability to distinguish between concern for second buyers and concern for sellers.

It should be emphasized that ratings *did not* persist between rounds. When subjects were randomly assigned to new groups at the beginning of each round, any ratings they received in previous rounds were not visible to the new group. This is essential to understanding the experiment, as it means that ratings were not accumulated throughout the course of each session, but existed only during the round in which they were given. Additionally, because subject roles were switched between rounds, the incentive to rate to influence a future partner was minimized.

Experimental Hypotheses

The first and most straightforward prediction to be tested is that a higher cost of rating will decrease the number of ratings, regardless of round type. This follows from the theoretical prediction that the likelihood that any arbitrary quality level is inside of the unrated range between $\underline{q}_i$ and $\overline{q}_i$ is decreasing in the size of that range.

Hypothesis 1 (Ratings Volume)**.** *The frequency of rating will be significantly higher when rating is free than when it is costly.*

Next, based on the theoretical predictions that $\frac{d\overline{q}_B}{dc}, \frac{d\overline{q}_S}{dc} > 0$ and $\frac{d\underline{q}_B}{dc}, \frac{d\underline{q}_S}{dc} < 0$, first buyers faced with a cost of rating should be more inclined to provide ratings for high or low quality

with the ability to send payoff-irrelevant messages after the proposers have made their offers. This suggests that subjects may simply wish to express their displeasure, even if it is not relevant to their earnings.

sellers than for moderate ones. This leads to our second hypothesis:

Hypothesis 2 (Polarization of ratings)**.** *Ratings will be more polarized in the Costly treatment than in the Free treatment. High and low quality sellers will receive a larger percentage of all ratings when rating is costly than when it is free.*

If first buyers provide ratings in buyer-fixed rounds, their actions must be an attempt to affect sellers in some way. The theory predicts that upon observing high quality ($q \geq \overline{q}_B$) they will give a positive rating to reward the seller. Likewise, if first buyers observe low quality ($q < \underline{q}_B$) they will give a negative rating to harm the seller in retaliation for offering low quality. These predictions lead to the next two hypotheses.

Hypothesis 3 (Altruism toward sellers)**.** *In buyer-fixed rounds, first buyers will give high quality sellers positive ratings, even when it is costly to do so.*

Hypothesis 4 (Retaliation toward sellers)**.** *In buyer-fixed rounds, first buyers will give low quality sellers negative ratings, even when it is costly to do so.*

Similarly, if first buyers rate sellers in seller-fixed rounds, they must be attempting to affect second buyers. In this case, a rating serves as an informative signal to second buyers and can be viewed as an altruistic act. [10]

Hypothesis 5 (Altruism toward buyers)**.** *In seller-fixed rounds first buyers will provide truthful ratings in order to aid other buyers, even when it is costly to do so.*

6 Results

Table **??** lists summary statistics for the experiment. Data is reported at the decision level, with each observation being a single decision made by one subject in one round. Because subjects interacted repeatedly and with randomly varying group membership across rounds, each decision cannot be treated as an independent observation. To account for this, much of the analysis below uses panel data regressions with standard errors clustered at the session level to account for unobserved heterogeneity. Results from several different specifications are reported for robustness, though they are qualitatively similar for each specification.

[10] It is possible that first buyers could provide intentionally misleading ratings specifically to harm second buyers. Indeed, there are a handful ($< 1\%$) of observations in the data that appear to be spiteful behavior toward second buyers.

Table 1: Summary statistics.

	Free **Buyer-Fixed**	**Free** **Seller-Fixed**	**Costly** **Buyer-Fixed**	**Costly** **Seller-Fixed**
Rating	3.194	3.216	2.225	2.543
	(1.469)	(1.477)	(1.711)	(1.725)
Prob. Rate	0.882	0.898	0.359	0.333
	(0.323)	(0.303)	(0.481)	(0.472)
Quality	5.299	5.087	3.173	3.493
	(2.688)	(2.681)	(2.925)	(2.950)

Standard deviations in parentheses.

6.1 Main Findings

Figure ?? below shows average ratings given in each treatment. Panel data regressions reported in columns 1 and 2 of table ?? show there is no significant difference in the value of ratings between buyer-fixed and seller-fixed rounds.[11] This means that, contingent upon giving a rating, subjects provide the same average ratings in both round types. Ratings are also increasing with cost, though this effect is quite small and only marginally significant.[12]

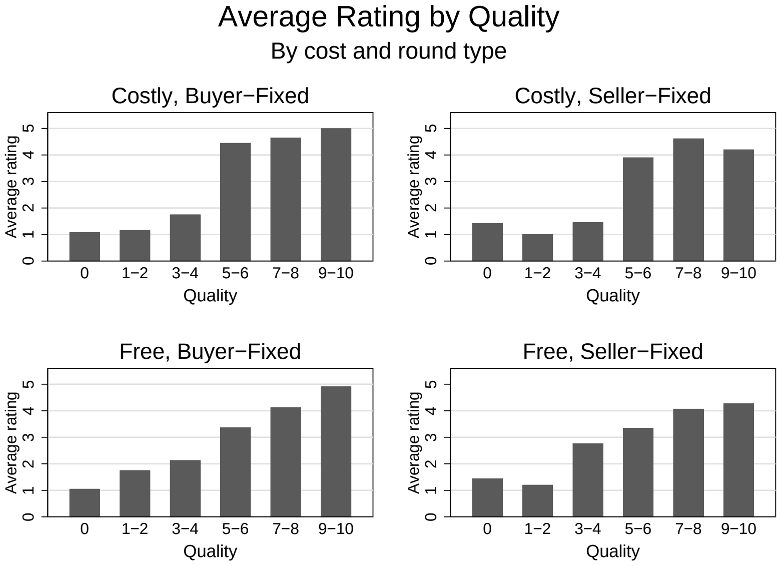

Figure 4: Average rating per quality bin, by cost and round type.

[11] All of the results in table 2 are robust to a variety of other specifications. In particular, the inclusion of interaction terms for the two treatment variables, cost and seller-fixed, provides nearly identical results.

[12] Cost is a dummy for the Costly treatment, and is not the actual $.25 cost of rating.

Table 2: Probability and choice of rating.

	Choice of rating (1-5)		Probability of rating			
	GLS		GLS		Probit	
	Random Effects	Fixed Effects	Random Effects	Fixed Effects	Random Effects	Fixed Effects
	(1)	(2)	(3)	(4)	(5)	(6)
period	0.00844	0.00882	0.00212	0.00213	0.00194	0.00211
	(0.00740)	(0.00762)	(0.00385)	(0.00384)	(0.00327)	(0.00332)
cost	0.136*	–	-0.569***	–	-0.538***	–
	(0.0821)		(0.0501)		(0.0203)	
seller-fixed	-0.00497	-0.00897	-0.0109	-0.0149	-0.00625	-0.0102
	(0.0714)	(0.0704)	(0.0205)	(0.0202)	(0.0194)	(0.0198)
quality	0.418***	0.426***	0.0117	0.00995	0.0151	0.0134
	(0.0214)	(0.0236)	(0.0103)	(0.00964)	(0.00952)	(0.00932)
|quality - 5|	-0.0177	-0.0189	0.0671***	0.0706***	0.0726***	0.0750***
	(0.0397)	(0.0408)	(0.0165)	(0.0165)	(0.0128)	(0.0125)
constant	0.973***	0.900***	0.668***	0.214**	–	–
	(0.224)	(0.183)	(0.108)	(0.0837)		
N	550	550	960	960	960	960

Panel regressions with standard errors clustered by session. Reported probit coefficients are marginal effects and thus do not include constants. Cost did not vary within sessions and thus is dropped from the session-fixed effects specifications.

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

Behavior is most interesting when we examine the binary decision of whether to rate, rather than the ratings themselves. Figure **??** shows the probability of rating by treatment. The results in columns (3) and (5) of Table **??** show that the cost of rating has a large and highly significant effect on the frequency of rating, with first buyers being more likely to rate in the Free treatment than the Costly treatment.

Finding 1. *The volume of ratings is significantly higher in the Free treatment than in the Costly treatment.*

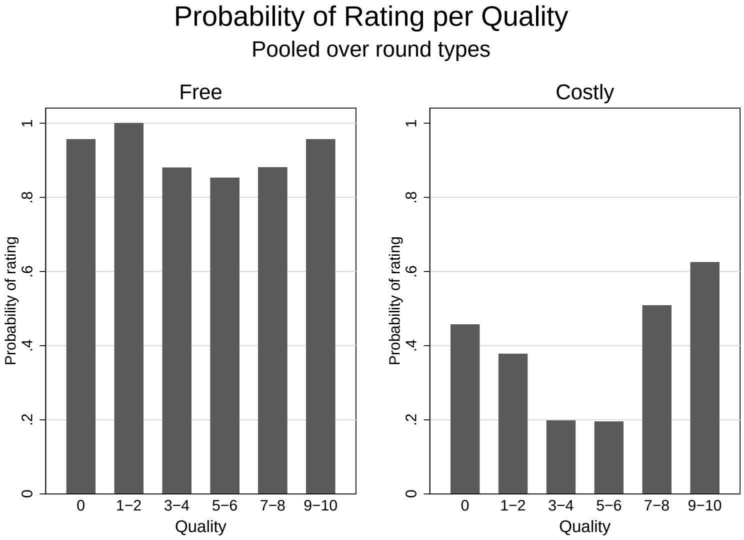

Figure 5: Probability first buyer rated, by quality and cost.

While it is not surprising that fewer ratings are given in the Costly treatment, the magnitude of the difference is striking. Removing a cost of only \$0.25, or approximately 2.3% of the average subject payment, leads to a more than 50 percentage point increase in the frequency of rating. Additionally, the change in frequency is not uniform across qualities. The |quality - 5| term shows that the probability of rating increases as quality becomes more extreme. In other words, high and low quality sellers are more likely to be rated than those of moderate quality. This shows us that polarization exists in the data, but does not explain its source. We next check how polarization varies between the Costly and Free treatments.

To examine this difference, we first define an extreme quality as any quality less than 4 or greater than 6, and a moderate quality as being those between 4 and 6 inclusive.[13] We then calculate the mean behavior of each subject when faced with moderate and extreme qualities.

[13] This division was chosen as it divides the range of qualities into equal and symmetric ranges, however these findings are robust to alternate definitions that shift or broaden the set of "moderate" qualities.

Each observation thus contains a subject's mean rating frequency for rounds in which they faced moderate quality (F_m) and rounds in which they faced extreme quality (F_e). We then calculate the ratio of rating frequency for moderate qualities to extreme qualities, F_m/F_e for each subject and use Somers' D test, with errors clustered at the session level, to find that the ratio is significantly greater when rating is free ($F_m/F_e = .91$) than when it is costly ($F_m/F_e = .33$), ($p < .001$). A ratio of 1 would indicate an absence of polarization, while a ratio near zero shows high polarization. In other words, introducing a cost of rating leads to increased polarization, with subjects being less likely to rate moderate qualities, relative to extreme ones.

Finding 2. *Polarization is significantly greater in the Costly treatment than in the Free treatment.*

Finding **??** gives support to the polarization hypothesis, and provides a first glimpse into what may be causing the U-shaped distributions observed in online rating data. It shows that, in the face of a small cost of rating, people are more willing to rate when they have either a very positive or very negative experience relative to a more moderate one. Finding **??** also lends support to the idea that raters rate to influence buyers and sellers, and not just out of a "joy of rating" or subject confusion. If subjects enjoyed rating for its own sake, or were confused as to the structure of the experiment, we would expect that their ratings would be uniform across qualities in the Costly treatment. Given that raters are highly responsive to quality in the Costly treatment, it must be the case that the benefit from rating is somehow derived from seller quality.

Next we examine whether subjects are more likely to rate in either buyer-fixed or seller-fixed rounds. Figure **??** shows the probability of rating different qualities for each treatment. The results in columns 3-6 of Table **??** show that there is no significant difference in the likelihood of rating between seller-fixed and buyer-fixed rounds. We also see that first buyers are no more likely to rate high quality sellers than low quality ones. Because ratings are given for both high and low quality sellers in each round type, this suggests that raters are driven to rate by altruism toward buyers and sellers, as well as revenge against sellers. These observations together give support to hypotheses 3-5, showing that raters are motivated by concern for both buyers and sellers.

6.2 Additional Observations

While the mean frequency of rating does not vary by round type, the level of polarization does. Using the same approach as before, we can examine the ratio of rating frequencies, to

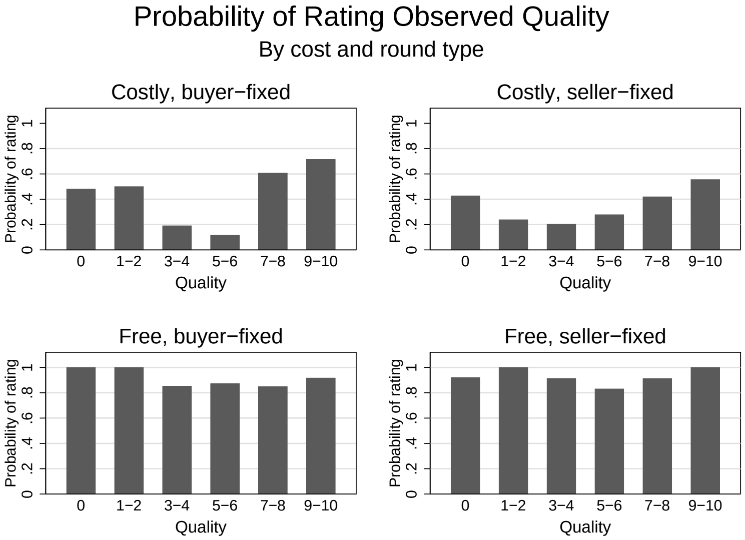

Figure 6: Probability first buyer rated, by quality, cost and round type.

see if raters are equally polarized in both round types. Using a Somer's D test with session-clustered standard errors shows that the ratio of moderate to extreme rating probabilities (F_m/F_e) is not significantly different between buyer-fixed and seller-fixed rounds, when rating is free ($p = .49$,). In the Costly treatment, however, the ratio is significantly greater in the seller-fixed treatment than in the buyer-fixed treatment ($p < .01$). In other words, in the Costly treatment we see more polarization in buyer-fixed rounds than in seller-fixed ones.

Finding 3. *When rating is costly, there is greater polarization of ratings in buyer-fixed rounds than seller-fixed rounds.*

Intuitively, finding **??** says that first buyers more often rate sellers of moderate quality when their goal is to affect other buyers rather than sellers. This is consistent with the idea of trying to provide useful information to other buyers, but only trying to cause or prevent a sale for sellers. This is important, as it says that ratings given with the goal of influencing sellers may be subject to greater polarization and therefore potentially greater bias in mean ratings. It also suggests that it may be more advantageous for designers of ratings system to focus raters on the impact their ratings have on other consumers, and to avoid the more strategic thinking that goes into trying to influence sellers.

The data on the probability of rating also describes the level of concern first buyers show for sellers and second buyers. For example, if we observe a rating given for quality $q < R$ in a second buyer-fixed round, we can infer that $q < \underline{q}_S$. This approach can only give us rough estimates, however, as we do not observe any actual values of R or beliefs about q' held by subjects. For this coarse estimate, we will assume $R = 5.14$ and $q' = 3.36$. Post-experimental questionnaires found the mean quality that subjects believed to be "fair" was 5.14, while $q' = 3.36$ was the mean quality offered by all sellers in the costly treatment.

Given our assumption that $R = 5.14$, ratings seen in the 5-6 quality bin in the Buyer-fixed treatment indicate the highest willingness to rate, as they observe a quality close to R. This is very close to "fair" or neutral treatment according to the theory, and the benefit from rewarding or punishing the seller is thus small, requiring a larger α to make rating worthwhile.

For this analysis, we use each subject's average probability of rating as the unit of observation. This is done to avoid overweighting the decisions of subjects who, due to random assignment, happened to be in the role of first buyer more often than others.

The 5-6 quality bin was rated only 12.4% of the time in buyer-fixed rounds. If this 12.4% is willing to rate the fairest (most neutral) possible quality, they should also be willing to rate any other, more extreme quality. We can then assume that 12.4% of the population will always rate, regardless of the quality offered by sellers. The 9-10 bin was rated 66.7% of the time, meaning that $100\% - 66.7\% = 33.3\%$ of the population appears unwilling to rate even the most extreme quality in the Buyer-fixed treatment. We then have that 100% - 33.3% -12.4% = 54.3% of the population that is willing to rate extreme qualities, but is unwilling to rate the most fair qualities.

Table 3: Types of Raters.

	Never Rate	**Sometimes Rate**	**Always Rate**
Buyer-Fixed	33.3%	54.3%	12.4%
Seller-Fixed	44.4%	34.2%	21.4%

We can apply a nearly identical approach to the Seller-fixed treatment. Our estimate of $q' = 3.36$ lies in the 3-4 quality bin, meaning that the observed quality is closest to the expected quality of other the other seller. We would expect those willing to rate the 3-4 bin to also be willing to rate more extreme qualities. The 3-4 bin was rated 21.4% of the time in seller-fixed rounds. The 9-10 bin was most frequently rated, at 55.6%, meaning that $100\% - 55.6\% = 44.4\%$ are unwilling to rate any quality, and, $100\% - 44.4\% - 21.4\% = 34.2\%$ of the population is willing to rate the most extreme quality, but will not rate a seller of the average quality observed in the experiment. It should be noted again that these are only very rough estimates, as we do not directly observe R or beliefs about q' in the experiment. This question of consistency of rating over time could be nicely addressed through observational studies of online rating behavior.

In addition to examining why ratings are given, it is also important to know the effect

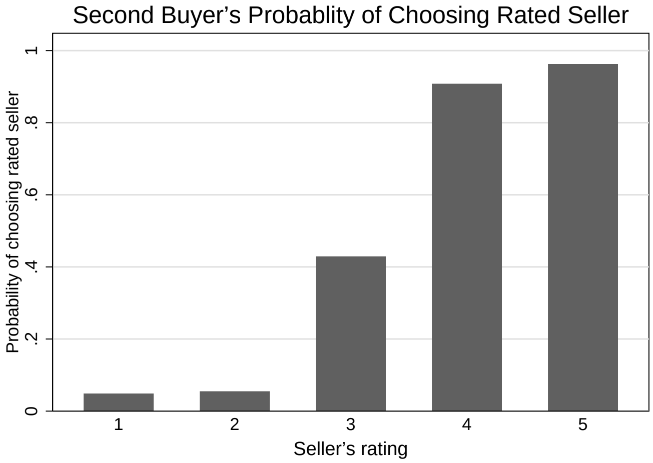

Figure 7: Probability second buyer chooses the rated seller, by rating

of ratings on market outcomes. Because rating is assumed to be an attempt to affect sellers and second buyers, it is reasonable to check what impact ratings have on behavior for each of those roles. Figure **??** shows that second buyers are heavily affected by the recommendations of first buyers. More than 95% avoid sellers with ratings of 1 or 2, and a similar percentage (94%) choose sellers with ratings of 4 or 5. Slightly less than half (43%) of second buyers choose a seller with a rating of 3, suggesting that buyers are essentially indifferent when faced with a middling rating. There is no significant difference in behavior between the Costly and Free treatments.

What impact does the cost of rating have on seller behavior? Regression results in column (1) of Table **??** show that the cost of rating has a large and highly significant effect on seller quality.

Finding 4. *Sellers offer significantly higher quality levels in the Free treatment than in the Costly treatment.*

Finding **??** indicates that the cost of rating is a significant factor in a seller's decision of what quality to provide to buyers. Quality rises from 2.96 in the Costly treatment to 4.86 in the Free treatment, an increase of 64%. This difference is especially striking when comparing the distributions of quality by treatment, as seen in Figure **??**. While 34.3% of sellers offer a quality of 0 in the Costly treatment, about one third as many, 12.4%, do so in the Free treatment. While the results also show a modest increase in quality over time, the difference between the Free and Costly treatments exists even in the first round.

Table 4: Quality and welfare regressions.

	Seller Quality (1)	Seller Welfare (2)	1st Buyer Welfare (3)	2nd Buyer Welfare (4)
period	0.0364**	-0.0267***	0.0310**	0.0544***
	(0.0168)	(0.00928)	(0.0136)	(0.0207)
cost	-1.894***	1.041***	-1.991***	-1.880***
	(0.472)	(0.266)	(0.484)	(0.488)
seller-fixed	0.0124	0.0114	0.0483	-0.0689
	(0.0925)	(0.0707)	(0.110)	(0.153)
constant	4.858***	7.554***	5.541***	5.807***
	(0.307)	(0.183)	(0.322)	(0.361)
N	1920	1920	960	1920

Panel data GLS regressions with standard errors clustered by session reported in parentheses.

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

This difference can be explained by sellers correctly anticipating the frequency of rating in each treatment. When rating is free, sellers anticipate that they are more likely to be rated when they offer low qualities and thus offer higher qualities to avoid a negative rating. When rating is costly, they know that it is relatively more likely that they will be able to offer low qualities and escape without a rating. A higher cost of rating thus results in lower quality being offered by sellers.

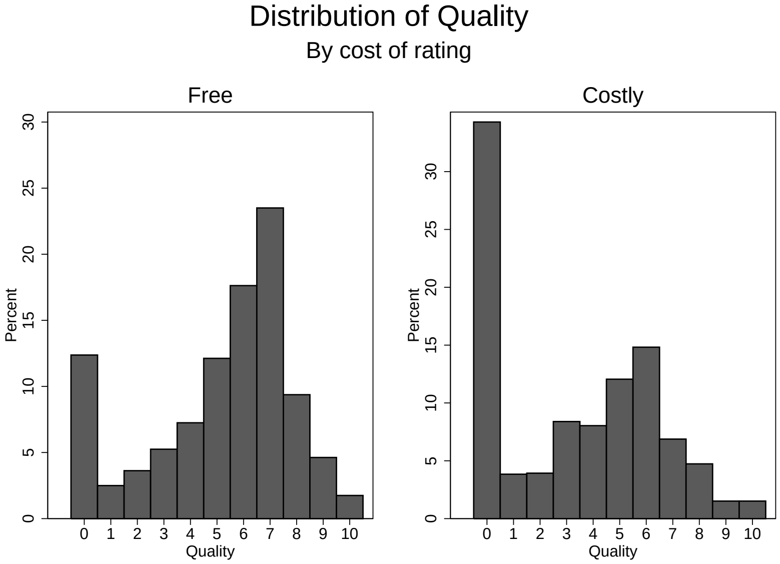

Figure 8: Distribution of qualities offered in the Free and Costly treatments.

Lastly, we want to ask how the cost of rating affects the welfare of buyers and sellers. Regression results in columns 2-4 of Table **??** show that cost has a significant effect on the adjusted earnings for each of the roles in the experiment.[14] Sellers' earnings are higher in the less frequently rated Costly environment, while first and second buyers' earnings are higher in the Free treatment. Notice that the nearly $2.00 increase in first buyer earnings in the Free treatment is much larger than the direct benefit of not paying the $0.25 cost of rating. We also see a small but significant time trend, with sellers being worse off over time while buyers are better. This mirrors the increase in seller quality over time, though it does suggest that sellers are actually worse off as they offer higher quality.

[14] Earnings reported are based on dollar amounts subject would receive if neither side of the market was fixed. Including the actual fixed payments skews the average earnings of sellers and second buyers towards the fixed earnings of $6.00.

7 Discussion

What can we learn from these findings, and how can they be applied? Understanding why people rate may help to improve the design of future rating systems. Different systems may affect ratings' ability to accurately reflect product quality. For example, consider a product which is of acceptable quality but has some small probability of failure. In the face of even a small cost, consumers who receive a functional, though unremarkable, product would be unlikely to provide a rating. However, the small number of consumers who do have a bad experience would be likely to provide negative ratings for the product. This would lead the product to have inaccurately poor ratings. As a simple example, consider a product which is generally of moderate quality, but occasionally fails utterly. If 80% of consumers receive a product of quality 5 and 20% receive quality 0, the average rating in the presence of a cost will be 3.07, compared with 3.6 when rating is free.

Designers of rating systems should pay special attention to minimizing any costs which may discourage consumers from providing ratings. Given that the very act of providing a rating may be burdensome, designers may want to provide small incentives to buyers for rating products. A small discount on future purchases, as in Avery et al. (1999) and Li and Xiao (2010), could be all that is necessary to offset the cost of rating.[15]

Another possible solution to this problem is to provide information about the total number of products sold. A consumer who is aware of this bias in ratings could correct for the distortion if she was aware of what proportion of people had purchased the product but left no rating. Providing this information could also help to increase sales via observational learning, as in Cai et al. (2009).

Designers should also be mindful of how they frame their requests for users to give ratings. They may receive different ratings if they focus consumers' attention on the seller or on future buyers. Requests that emphasize helping other buyers may produce ratings driven more by comparisons with other products or sellers, whereas requests that focus on the sellers will have a greater focus on fairness. It is not clear if a rating based on perceived fairness or relative quality is more desirable in general, as there are likely scenarios in which either bias is preferred.

[15] It should be noted, however, that this approach has the potential to create new incentive problems. Consumers who are motivated solely by a monetary reward for leaving *any* rating may not be concerned with the accuracy of their ratings.

8 Conclusion

This paper examines the factors that influence consumers' decisions to rate products online. Evidence from a laboratory experiment shows that consumers are motivated to rate both by a concern for punishing or rewarding sellers and by a desire to inform future buyers. Introducing even a small cost of rating has a large effect on rating behavior, leading to fewer and more polarized ratings. One implication of this finding is that any cost of rating, even a small and implicit one, may cause a "blind spot" in ratings distributions. This can cause inaccurate average ratings for products of variable quality, especially those whose quality distributions are asymmetric. This polarization may also be more intense when ratings are focused on sellers than when they are focused on buyers.

Sellers are responsive to buyers' cost of providing ratings, and adjust their quality accordingly. A small decrease in the cost of rating causes a significant increase in the level of quality offered by sellers. This contributes to existing evidence, such as Bolton et al. (2004), that suggests consumer-generated ratings systems may significantly increase consumer welfare. This finding further demonstrates that the cost of rating should be a major concern for designers of rating systems.

The experimental design introduced in this paper may have significant future applications, both for ratings research and elsewhere. Probabilistically deactivating payoffs as described here can disentangle other settings involving multiple simultaneous motivations. In addition to studying a variety of other ratings environments, such as student evaluations and expert reviews, the design can be adapted to disentangle motivations in other three-party interactions. For example, it can be applied to study educators' incentives to provide grades to their students. Are grades given to aid employers in choosing the best candidates, or to help students land the best jobs? It may be used to understand policing behavior, separating a desire to protect potential victims from a desire to punish perpetrators. It can also apply to understand why aid is given to person or country in conflict: Is it to help the recipient, or to hurt the recipient's foe? In short, any similar instance in which one party is acting to influence the interaction of two other parties may be amenable to this design.

This paper's findings cannot exclude other possible motivations for the voluntary provision of ratings. For example, the first Harry Potter book has more than 8,000 ratings on Amazon.com. The likelihood that the 8,000th rating will have any effect on either buyers or sellers is very low. A rational consumer, faced with $7,999$ previous ratings, would be unlikely to rate if they were motivated solely by a desire to influence buyers and sellers. Such a buyer may still rate if they enjoy the act of rating, or if they enjoy the thought that their rating might be seen by others. They may also rate if they misperceive the likelihood of their rating being impactful. While these alternate motivations are not examined in this work, it is likely that they are the source of at least some of the ratings found on the internet.

There are several directions for future research. One important feature of online ratings that has been intentionally removed from this setting is the accumulation of ratings over time. Because only one person can be the first rater for each product, in practice most ratings are given in the shadow of many previous buyers' opinions. Given a series of pre-existing ratings by other raters, do consumers provide their honest opinion of a product or do they attempt to adjust the mean rating toward what they feel is the correct value?

It would be valuable to see portions of this experiment replicated in a more natural environment. While the laboratory gives us unrivaled control over experimental conditions, it would be useful to document the behavior described in this paper "in the wild." In particular, it would be interesting to see how the distribution of ratings for a real product varies with implicit and explicit costs of rating.

Acknowledgements

I am grateful to John Duffy for guidance, support and encouragement. I also thank Sourav Bhattacharya, Andreas Blume, Oliver Board, Tim Hister, Wooyoung Lim, Lise Vesterlund, Roberto Weber and seminar and conference participants at the University of Pittsburgh, Lehigh University, 2009 International and US Economic Science Association meetings, and 2009 Western Economic Association International meetings. I gratefully acknowledge the University of Pittsburgh for financial support. All errors are my own.

References

(1) Avery, Christopher, Paul Resnick and Richard Zeckhauser. 1999. "The Market for Evaluations" *American Economic Review,* 89(3): 564-584.

(2) Bolton, Gary, Elana Katok and Axel Ockenfels. 2004. "How Effective are Electronic Reputation Mechanisms? An Experimental Investigation." *Management Science,* 50(11): 1587-1602.

(3) Cai, Hongbin, Yuyu Chen and Hanming Fang. 2009. "Observational Learning: Evidence from a Randomized Natural Field Experiment." *American Economic Review,* 99(3): 864-82.

(4) Chen, Yan, Max Harper, Joseph Konstan and Sherry Xin Li. 2010. "Social Comparisons and Contributions to Online Communities: A Field Experiment on MovieLens." *American Economic Review,* 100(4): 1358-98.

(5) Chevalier, Judith and Dina Mayzlin 2006. "The Effect of Word of Mouth on Sales: Online Book Reviews." *Journal of Marketing Research* 43(3): 345-354.

(6) Dellarocas, Chrysanthos, Neveen Farag Awad and Xiaoquan (Michael) Zhang. 2004. "Exploring the Value of Online Reviews to Organizations: Implications for Revenue Forecasting and Planning."*Planning Proceedings of the 25nd International Conference on Information Systems (ICIS)*, Washington, D.C

(7) Duan, Wenjing, Bin Gu and Andrew B. Whinston. 2008. "Do Online Reviews Matter? – An Empirical Investigation of Panel Data." *Decision Support Systems*, 45(4): 1007-1016.

(8) Fehr, Ernst and Simon Gächter. 2000. "Cooperation and Punishment in Public Goods Experiments." *American Economic Review,* 90(4): 980–994.

(9) Fischbacher, Urs. 2007. "z-Tree: Zurich Toolbox for Ready-made Economic Experiments." *Experimental Economics,* 10(2): 171-178.

(10) Houser, Daniel and John Wooders. 2005. "Reputation in Auctions: Theory, and Evidence from eBay." *Journal of Economics and Management Strategy*, 15(2): 353–370.

(11) Hu, Nan, Jie Zhang and Paul A. Pavlou. 2009. "Overcoming the J-shaped distribution of product reviews." *Communications of the ACM - A View of Parallel Computing*, 52(10) 144–147.

(12) Levine, David K. 1998. "Modeling Altruism and Spitefulness in Experiments." *Review of Economic Dynamics*, 1 (3): 593–622.

(13) Li, Lingfang(Ivy) and Erte Xiao. 2010. "Money Talks? An Experimental Study of Rebate in Reputation System Design" forthcoming in *Management Science.*

(14) Servátka, Maroš. 2009. "Separating Reputation, Social Influence, and Identification Effects in a Dictator Game." *European Economic Review*, 53(2): 197–209.

(15) Wang, Zhongmin. 2010. "Anonymity, Social Image, and the Competition for Volunteers: A Case Study of the Online Market for Reviews." *The B.E. Journal of Economic Analysis & Policy*, 10(1), Article 44.

(16) Erte, Xiao and Daniel Houser. 2008. “Emotion Expression in Human Punishment Behavior.” *Proceedings of the National Academy of Sciences of the United States of America*, 102(20): 7398–7401.

(17) Li, Xinxin and Lorin Hitt. 2008. “Self Selection and Information Role of Online Product Reviews.” *Information Systems Research*, 19(4): 456–474.