Ratings with Heterogeneous Preferences

Jonathan Lafky
Carleton College

Robin Ng
University of Mannheim

August 2024

Abstract

We examine how product ratings are interpreted in the presence of heterogeneous preferences among both raters and consumers. Raters with altruistic motives should rate for the benefit of future consumers, however an ambiguity arises when preferences are heterogeneous. Multiple equilibria exist in which ratings may reflect the preferences of raters or the preferences of future consumers. In an online experiment, we examine how ratings are selected by raters and interpreted by consumers, and how information about rater preferences or product attributes can influence equilibrium selection. We show how both raters and consumers update their evaluation of what a rating represents in each environment, doing so in similar ways.

JEL: C91, D64, D83, L86
Keywords: Ratings and Reviews, Altruism

1 Introduction

We rely on the experiences of others to make many decisions. Increasingly, we learn about those experiences via consumer-based ratings, which influence a wide variety of decisions such as where we eat, what medical care we receive and which toys we buy our children. Consumers use ratings to make more informed consumption decisions, but only if they are able to correctly interpret what a rating represents. This is a relatively simple task if all consumers share similar preferences; however, in many cases different consumers can have very different tastes for the same product. Heterogeneous preferences create an important ambiguity: Whose preference does a rating reflect?

One illustration of this ambiguity can be seen in consumer ratings for hotels, a market where consumers can have starkly different preferences from one another. Some consumers prioritize location, such as ease of access from the airport or proximity to conference venues. Others are less concerned with location and focus instead on the experience within the hotel, such as the cleanliness of their room. Given these diverse preferences, how does a consumer rate the hotel at the end of their stay? How much does a traveler concerned primarily with location also account for room cleanliness when rating? In general, do raters rate based on their own preferences, or do they consider the preferences of others?

Ratings are used in a variety of contexts that provide consumers with other sources of information about products or about raters themselves. In many settings consumers can directly observe some of a product's attributes prior to purchase, decreasing not only the consumer's uncertainty of the product's quality, but also potentially diminishing the ambiguity of what a rating represents. Returning to the case of hotels, location is an easily verifiable attribute, while cleanliness cannot be observed in advance. Since consumers can already observe a hotel's location, there is no need for raters to repeat that information, allowing them instead to focus on only the unknown attribute, cleanliness. Rating over just the unknown attribute makes consumers better off, but only if raters and consumers can coordinate on that convention. Although evaluating the unknown attribute is most beneficial to consumers, it may nonetheless be more natural to evaluate the product as a whole, and it is not clear which of these norms occurs in practice.

In addition to learning specific product attributes, consumers often have, or are able to infer, information about a rater's preferences. For example, many written reviews accompanying ratings will describe specific attributes the rater cares about, such as "Hotel staff were very responsive." or "I chose this hotel because of the view." On many ratings platforms it is also possible for consumers to view a rater's profile, allowing them to see how the rater has evaluated other products in the past. Regardless of the source, shared information about rater preferences may decrease the ambiguity of what a rating represents, as a rater who is known to prioritize one attribute (e.g. location) may be more likely to rate based on that attribute.

We analyze these ratings environments via a simple theoretical model and online experiment, demonstrating how ratings are created and interpreted in the presence of heterogeneous preferences.[1] We first construct a baseline setting where consumers observe only ratings, and then examine how rater and consumer interpretations of ratings change as we provide information about rater preferences and product attributes. Our key result, which applies across all treatments, is that raters heavily favor their own preferences when leaving a rating, and consumer interpretations of ratings are largely consistent with raters. In other words, the ambiguity arising in the presence of heterogeneous preferences is resolved by raters and consumers both focusing on rater's preferences.

When consumers are informed of rater's preferences alongside each rating, consumer sensitivity to ratings hinges upon having the same preference as the rater; consumers who share preferences with the rater become more sensitive to ratings, while those with different preferences become less sensitive. When we instead allow consumers to directly observe one of the product's attributes, rater behavior does not change. Raters continue to describe the entire product, contrary to the prediction that mutually known attributes should be ignored.

The rest of the paper is structured as follows. Section 2 discusses the related literature and our contribution to it. We describe our theoretical model in Section 3, experimental design in Section 4 and hypotheses in Section 5. In Section 6, we discuss the results of our experiment, and Section 7 concludes.

2 Related Literature

Much of the literature on online ratings and reviews focuses on implications for firm pricing and consumer purchasing decisions (Cabral and Hortacsu, 2010; Dellarocas et al., 2007; Li et al., 2020; Luca and Reshef, 2021; Mayzlin et al., 2014). Evidence from online marketplaces (Cai et al., 2014; Jin and Kato, 2006) and in the laboratory (Bolton et al., 2004; Halliday and Lafky, 2019) have shown that ratings increases trust between sellers and buyers.

An extensive empirical literature has examined when and how raters choose to rate. Unsurprisingly, quality has been shown to be an important driver of ratings (Hui et al., 2023; Li et al., 2020; Proserpio et al., 2018; Zhang et al., 2012). This is the focus of most of the theoretical literature which abstracts from the decision to rate by taking ratings

[1]For the sake of simplicity, we model environments in which all consumers have the same ordinal preferences within each of the product attributes, but the relative weights between attributes vary between consumers.

as a given and modelling ratings as increasing in quality (Cabral, 2000; Tadelis, 1999). Other factors such as reference prices (Gesche, 2022), hidden fees (Chiles, 2021), social distance (Masterov et al., 2015), and power distance (Gao et al., 2018) have also been shown to influence rating decisions.

A growing empirical literature suggests social preferences motivate the decision to provide ratings (Bolton et al., 2013; Chakraborty et al., 2022; Chen et al., 2010; Fradkin et al., 2021; Qiao et al., 2020). Some experiments have more precisely shown that altruism is one such social preference (Hoyer and van Straaten, 2022; Lafky, 2014). Since ratings are a public good, these experiments connect well to the broader literature suggesting altruism as an important motivation for public goods contributions (see Bowles and Polania-Reyes (2012) for a survey).

We study a communications game where raters and consumers have identical objective functions, and potentially different preferences. Since raters do not face any explicit cost or benefit from rating, our environment is similar to cheap-talk (Crawford and Sobel, 1982). However, because raters and consumers have identical objective functions, the traditional tension found in cheap-talk environments is absent from our setting. Further, by allowing for heterogeneous preferences, we contribute to the ongoing discussion on how preference similarities can affect trust between senders and receivers (Connors et al., 2011; Woodside and Davenport Jr, 1974).

Identical product listings have been shown to obtain different ratings across multiple websites (Chevalier and Mayzlin, 2006; Schneider et al., 2021; Zhang et al., 2012), which can pose a problem when comparing products across websites. Chevalier and Mayzlin (2006) attribute this difference to buyer-self-selection onto platforms, Zhang et al. (2012) suggest cultural differences affect how people rate, while Schneider et al. (2021) argue multi-dimensional ratings affect the choice of ratings. We explore an alternative explanation, that contextual information may influence the interpretation of ratings.

3 Theoretical Model

3.1 Environment

We provide a simple two-player, two-stage sequential model in which a consumer leaving ratings derives utility both from their own consumption and from the surplus of a future consumer. To fix terminology, we label a consumer who generates ratings the "rater" and a future consumer that uses ratings the "consumer."

Product. We assume a product with two dimensions of quality described by the attributes, X and Y, with realizations x and y respectively. The quality levels are independently drawn from a commonly known distribution, and the price of the product, p, is fixed.

Consumption Utility. Both the rater and consumer have the same ordinal preferences within each product attribute.[2] However, the relative strength of their preferences across attributes may differ and depend on two independently drawn weights, α_i and β_i, where α_i, $\beta_i \in [0, 1]$, and $i \in \{r, c\}$ represents the rater and consumer. These weights correspond to preferences for the dimensions X and Y respectively, with consumption utility given by $u_i = \alpha_i x + \beta_i y - p$.

For the consumer, consumption of the product is the only source of utility in the model, so their utility is simply $u_c = \alpha_c x + \beta_c y - p$ if they choose to consume, and 0 otherwise. In expectation, this utility depends on the information available to the consumer, including any rating sent by the rater. Hence, the consumer chooses to purchase when $E_c[u_c|R, I] \geq 0$, where $R \in \{R_p, R_n, R_\emptyset\}$ represents a positive, negative, or no rating respectively, and I captures all other information available to them.

Rater Utility. In addition to their own consumption, the rater also has concern for the consumer's welfare, represented by a weight $\kappa \geq 0$. Although the rater wants the consumer to make better choices, the act of rating may be burdensome due to the time and effort required. We model this as a cost of rating, $e > 0$. Therefore, the rater's utility is:

$$u_r = \alpha_r x + \beta_r y - p + \mathbb{1}_{\text{consumers buy}} \cdot \kappa E_r[u_c|x, y] - \mathbb{1}_{\text{rater rates}} \cdot e.$$

This expression captures that a rater's utility takes two parts. First, they receive some consumption utility based on their own preferences and the product's attributes. Second, they may receive additional utility based on the anticipated value of the product to the consumer. If the rater anticipates that the consumer buys, this may either be positive or negative. However, if the rater anticipates the consumer does not purchase this additional utility is zero. In other words, the additional utility comes from a rater's expectation consumers would buy the product based on the available information to the consumer, $E_r[u_c|R, I] \geq 0$.

While prices are included in the rater utility above for the sake of completeness, in equilibrium, the price a rater pays does not influence their rating decision. Going forward we consider the price a rater pays as a sunk cost and omit it from the rater's decision making process.

[2]Higher values of x and y positively affect their utility, but may not do so equally.

Interpreting Ratings. We suppose that there is an equilibrium interpretation of ratings. In particular,

$$R = \begin{cases} R_p & \text{if } F(x,y) > \overline{w} \\ R_n & \text{if } F(x,y) < \underline{w} \\ R_\emptyset & \text{otherwise} \end{cases} \tag{1}$$

where $F(x,y)$ is a combination of the values x and y such that $\partial F(x,y)/\partial x \geq 0$ and $\partial F(x,y)/\partial y \geq 0$, and $\overline{w}$ and $\underline{w}$ are cutoffs representing relatively "good" or "bad" quality, respectively.

3.2 Theoretical Findings

We briefly summarize our theoretical findings here, the details of which can be found in Appendix A.

Observation 1. *There are multiple equilibrium interpretations for ratings.*

Observation 1 reflects how any convention in which raters and consumers agree upon the mapping of x and y values into a rating is a possible equilibrium. In other words, ratings may reflect the preferences of raters, those of consumers, or a combination thereof. Although theory admits a range of equilibria, some arise more naturally in specific environments. We consider how the incentives of raters and the environment through which ratings are transmitted can aid in equilibrium selection.

In every environment, raters prefer to provide ratings which are as useful as possible to consumers. When ratings are the only information available to consumers, ratings are most useful when they reflect the preferences of the population instead of just the rater. Hence, we anticipate that raters do not rate according to their own preferences, and select ratings which reflect the preferences of the broader population.

Observation 2. *When a rating is the only information transmitted to consumers, ratings should reflect the average consumer preference, independent of the rater's own preferences.*

In environments where raters' preferences are common knowledge, that information can potentially resolve the ambiguity of multiple equilibria by serving as a focal point: raters and consumers sharing knowledge of rater preferences makes it more natural for raters to simply reflect their own experience when evaluating products. We summarize this as our next observation.

Observation 3. *Common knowledge of rater preferences creates a focal point for ratings. Raters and consumers are then more likely to adopt an equilibrium based on those preferences.*

In instances where some attributes of a product's quality can be directly observed by consumers, it is not informative to incorporate those attributes when rating. Evaluating commonly known attributes does not improve consumers' understanding of the product, meaning that raters who want to leave the most informative ratings should ignore attributes of a product that can be independently observed by consumers. As an example, if the realization of attribute x is known, a rating should only take into account the attribute y, that is $\partial F(x, y)/\partial x = 0$ and $\partial F(x, y)/\partial y > 0$.

Observation 4. *When some product attributes are common knowledge, ratings are most informative when raters rate based on the unknown attributes.*

We next describe our experiment, designed to show how raters and consumers resolve the equilibrium selection problem arising from the presence of heterogeneous preferences.

4 Experimental Design

Our experiment comprises two parts, and was conducted using the oTree software (Chen et al., 2016) on Prolific. A total of 502 subjects were recruited from a sample of the US population. Each subject read a series of instructions, completed a comprehension quiz, and then played 20 rounds of the experiment. Subjects were paid a show-up fee of \$3.00 USD, and were able to earn bonus payments in tokens, with each token worth \$0.50 USD. The average subject took just above 13 minutes to complete the experiment and received a total payment (show up fee and bonus) of \$6.37 USD. There were minimal differences in the instructions between treatments, which are highlighted in Appendix B.

Each treatment took place in two phases. In the first phase, subjects were recruited to play the role of raters who rated randomly generated products. After collecting all first-phase data, a new group of subjects, excluding those who participated in the first phase, was recruited for the second phase to act as consumers who viewed ratings generated in the first phase. In each phase, subjects participated in 20 rounds of decision making. At the beginning of the 20 rounds, each subject was randomly assigned a preference for X (X-type) or Y (Y-type). This preference was fixed for all rounds.

Raters. In each of the 20 rounds, subjects in the rater role were asked to evaluate a two-dimensional product referred to in the instructions as a "prize." The product was composed of two values, X and Y, and both values were drawn from a discrete uniform distribution, $U\{1, 10\}$. Subjects who were assigned a preference for X valued the prize as $1 \times x + 0.1 \times y$, while subjects who were assigned a preference for Y valued the same prize as $0.1 \times x + 1 \times y$. Raters were informed of the consumer's task, and that consumers would also be randomly assigned to be either X- or Y-type. Raters were also told what information consumers would know prior to making their own decisions, which varied by treatment as described below.

Each round began with a prize being randomly generated for each rater. The rater learned the values of x and y, then rated the prize on a scale of 1 to 5, where 1 was stated to be the worst rating, and 5 the best. We use a 1-5 scale due to it being commonplace in the field and therefore natural for our subjects. This differs slightly from the binary ratings in our theory, however our predictions depend simply on the positive or negative signals conveyed by a rating and not upon the specific scale. Additionally, although our theory includes prices for raters, the price a rater pays does not affect their rating decision. To focus our experiment on how raters evaluate products, we abstract away from the rater's purchase decision and treat prices paid by raters as a sunk cost.

After rating the prize, raters were given the choice to share their rating with consumers for a small fee of 0.1 tokens ($0.05 USD). The fee reflects the opportunity cost of rating and ensures that ratings are only sent if raters have at least a minimal concern for consumers, thereby filtering out individuals who are indifferent between rating and not. Moreover, by imposing a cost rather than explicitly including the payoff of consumers in the rater's utility, any rating sent must be motivated by the intrinsic altruism of participants. At the end of the 20 rounds, one round was randomly selected and the value of the prize from that round, less any cost of sending a rating in that round, was paid to the rater.

Consumers. Like raters, each consumer was assigned to be X-type or Y-type, valuing the prize at $1 \times x + 0.1 \times y$ or $0.1 \times x + 1 \times y$ respectively. Consumers learned about the rater's task and that rater types were randomly assigned. Consumers were further split into two groups. The first group of consumers were shown only prizes for which raters had sent ratings. This group is the primary source of data for our analysis of consumer behavior. A second group of consumers saw only prizes for which ratings were not sent. This group was not used in our analysis, but was necessary to ensure raters knew unrated products would still be seen by consumers. Because our interest is in how consumers interpret ratings, we recruited twice as many subjects into the group with rated products than into the group with unrated products. Consumers knew in advance which group they belonged to, but they were not informed of the existence of the other group.

Each consumer played 20 rounds, and in each round was presented with a different prize. These prizes were sampled without replacement for each consumer, but with replacement across consumers. In other words, no consumer could draw the same prize more than once, however multiple consumers were able to draw the same prize.

Each consumer next indicated a willingness to pay (WTP) for that round's prize via a Becker-DeGroot-Marschak (BDM) Mechanism (Becker et al., 1964) similar to Healy (2018). Consumers were given a series of questions asking them to choose between the prize or an increasing number of tokens, and then indicated the question at which they would begin to choose tokens over the prize. The choices ranged between 0.1 and 11.0 tokens in increments of 0.1.

At the end of each round, consumers were told the realized values x and y. At the end of the 20 rounds, one round was randomly selected and their decision on one randomly selected question from that round's BDM was paid to the consumer.

Treatments. The experiment implements a 2×2 design, where information is varied over two dimensions. We first modify the information provided to the consumer about the rater's preferences. In Pref(erence) treatments, consumers learn the rater's type, and hence which of the two values the rater preferred (x or y), prior to stating their WTP. Conversely, in no-Pref treatments, they never learn the preferences of the rater.

We next modify the information provided to the consumer about the product's attributes. In Attr(ibute) treatments, consumers directly observe the realization of one attribute, x, prior to indicating their willingness to pay. Conversely, in no-Attr treatments, consumers do not directly observe any information about product quality. Note that revealing only x in Attr treatments is intentionally asymmetric, resulting in some consumers (X-types) having consistently better information than others (Y-types).

All combinations of our treatment variables gives us four treatments: (i) **None**, where consumers learn nothing about rater preferences or product quality; (ii) **Pref**, where consumers learn which of the two dimensions the rater preferred but nothing about product quality; (iii) **Attr**, where consumers do not learn rater preferences but do learn the value of x; (iv) **Both**, where consumers learn both rater preferences and the value x takes. The four treatments are summarized in Table 1.

		Product Information	
		No	Yes
Rater Preferences	No	(i) None	(iii) Attr
	Yes	(ii) Pref	(iv) Both

Table 1: Information available to consumers in each treatment.

5 Hypotheses

Our first three hypotheses focus on how raters choose to rate: either evaluating the product based on their own preferences or considering the preferences of others.

Our first prediction is that, when ratings are the only information available to consumers, raters provide ratings independent of their own preferences. This follows from Observation 2.[3]

[3]To connect this to our theory, suppose that prices are drawn from a commonly known distribution as reflected by choices within the BDM.

Hypothesis 1. *In the baseline None treatment, ratings are not sensitive to the preferences of the rater.*

Our second hypothesis addresses how information about rater preference affects the choice of rating. Although our theory allows for multiple equilibria, it is intuitive to think that consumers who learn rater preferences incorporate this information when interpreting ratings. Observation 3 suggests raters make their own preferences a focal point when choosing ratings. For example, if a rater reveals a preference for a comfortable bed, it is more likely for a consumer to interpret a rating as reflecting bed comfort, and less so room service.

Hypothesis 2. *Ratings are more sensitive to the rater's preferred attribute in the Pref treatments.*

Next, consider the environment where the quality of a product's attribute is public knowledge. Observation 4 suggests raters should not consider the known attribute when choosing a rating. This leads to our third hypothesis:

Hypothesis 3. *Ratings are not affected by the revealed attribute in the Attr treatments.*

Our last two hypotheses focus on how consumers interpreted ratings in each of the information environments, beginning with the Pref treatments where rater preferences are known. If consumers anticipate that ratings reflect the rater's preferred attribute, consumers who share the same preference are more likely to trust ratings. Conversely, consumers who do not have the same preference as the rater will be less influenced by ratings.

Hypothesis 4. *In the Pref treatments, consumer willingness to pay is more sensitive to ratings when consumers share the same preferences as the rater.*

Recall in the Attr treatments, the quality of a product's x attribute is publicly known. Since X-type consumers prefer this attribute, most of the value they receive from the product can be observed independent of a rating. Conversely, Y-type consumers observe little about their own value for the product beyond what they learn from the rating. As a result, compared to Y-type consumers, X-type consumers' willingness to pay is less sensitive to ratings. This leads to our final hypothesis.

Hypothesis 5. *In the Attr treatments, the willingness to pay of Y-type consumers is more sensitive to ratings than willingness to pay of X-type consumers.*

6 Results

Our analysis proceeds by first studying how raters select ratings for different products, then examining how consumers interpret those ratings. Table 2 provides summary statistics for the experiment, with mean values reported at the decision level. A total of 502 subjects each completed 20 rounds. For each treatment, 50 subjects were raters, 50 were consumers who saw ratings, and 25 were consumers who received no ratings.[4] Mean ratings were similar across all four treatments, ranging from 3.16 in Pref to 3.27 in Both. The frequency with which ratings were sent was low and variable across treatments, ranging from 20% in the Attr to 35% in Both. Better ratings were slightly more likely to be sent, as the mean sent rating ranged from 3.57 in Pref to 3.86 in None.

Although the treatments without ratings were only conducted to ensure that raters were not potentially misled, we do find that consumer willingness to pay was slightly higher with ratings than without. When consumers saw no ratings, their mean willingness to pay ranged from 49.42 in Both to 54.05 in Pref. Consumer mean willingness to pay was larger in all treatments where they saw ratings, ranging from 54.65 in None to 61.00 in Both. To understand how ratings are interpreted with heterogeneous preferences under different information environments, we restrict the remainder of our analysis of consumers to those who saw ratings.

	None	**Pref**	**Attr**	**Both**
Raters				
Rating	3.25	3.16	3.26	3.27
	(1.38)	(1.35)	(1.30)	(1.33)
Sent rating	3.87	3.57	3.59	3.84
	(1.21)	(1.41)	(1.45)	(1.27)
% sent	23	28	20	35
Subjects	51	51	50	50
Consumers (with ratings)				
WTP	54.65	54.39	60.40	61.00
	(29.54)	(31.77)	(32.54)	(27.96)
x	5.09	6.84	6.80	6.54
	(3.18)	(2.92)	(2.36)	(2.90)
y	6.73	5.76	5.71	6.06
	(2.68)	(2.75)	(2.71)	(2.47)
Subjects	50	50	50	50
Consumers (without ratings)				
WTP	52.46	54.05	52.49	49.42
	(28.33)	(27.90)	(32.41)	(30.50)
x	5.60	6.77	5.81	5.16
	(2.90)	(2.88)	(2.91)	(2.62)
y	5.65	5.84	5.50	5.33
	(2.78)	(2.83)	(2.76)	(2.51)
Subjects	25	25	25	25

Table 2: Summary statistics. Mean values with standard deviations in parenthesis.

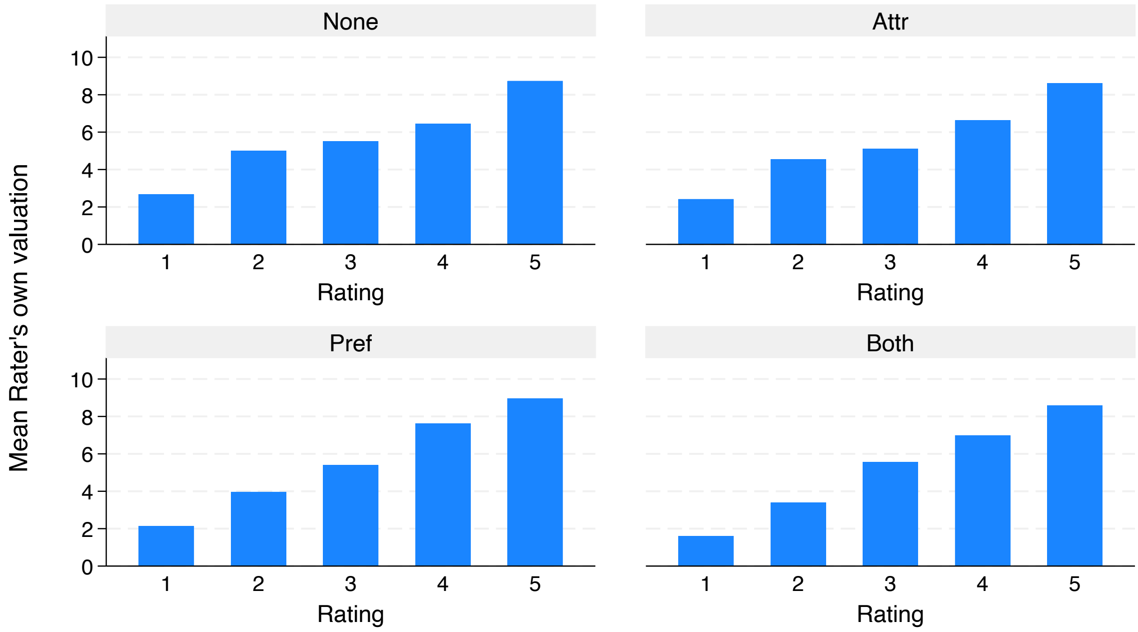

Figure 1: The average value of the product to the rater, for each sent rating.

6.1 Rater Behavior

We examine rater behavior first in terms of how raters selected their ratings and then whether they decided to send those ratings to consumers. Figure 1 shows a monotonic relationship between the rater's value of the product and the choice of rating, among those ratings that were sent to consumers.

Table 3 reports regression results on rater behavior. Column 1 shows the effects on rating choice in the full sample, while column 2 focuses on the sample restricted to sent ratings. The variable *ownvalue* describes the quality level for the rater's preferred attribute (x for X-type raters, y for Y-type raters) while the variable *othervalue* captures the rater's less-preferred attribute (y for X-type raters, x for Y-type raters). We include dummies *pref* and *attr* as treatment variables, where *pref* takes the value 1 in treatments when the preference of raters is public information, and *attr* takes the value 1 in treatments when the quality of attribute x was public information. We also interact *ownvalue* and *othervalue* with each of the treatment dummies to examine whether raters evaluated each of the attributes differently across treatments.

In all treatments, there is a large and highly significant effect of *ownvalue* on the choice of rating, alongside a much smaller effect of *othervalue*. The small effect of *othervalue* is only significant for the full sample of chosen ratings, and is insignificant when conditioning

[4]Due to a logistical mistake made during recruiting, 51 (rather than 50) subjects were recruited as raters in None and Pref.

upon only sent ratings. Together, these effects suggest that raters' decisions are primarily motivated by their preferred attribute (*ownvalue*) and show little concern for their less-preferred attribute (*othervalue*). In other words, raters rated almost exclusively based on their own preferences, rejecting Hypothesis 1. The interaction of *pref* with *ownvalue* and *othervalue* shows this effect is stronger in the Pref treatments, when raters knew that their own preferences were visible to consumers, consistent with Hypothesis 2.

The interaction of *attr* with *ownvalue* and *othervalue* leads to no statistically significant effects, at first seeming to suggest that raters did not rate differently between the Attr treatments and baseline. However, recall that our prediction is that ratings in the Attr treatments reflect the quality of the unknown product attribute, which is independent of the rater's preferences. To test our predictions for the Attr treatments, we need to examine the relationship between choice of rating and the x and y, rather than *ownvalue* and *othervalue*. Specifically, raters should ignore the commonly known x, and rate solely based on the unknown y.

Column 3 of Table 3 shows how the choice of rating depends upon the x and y with and without information about the attribute x. There are no significant effects from interacting *Attr* with x or y, in contrast to our predictions from Hypothesis 3. Raters appear to be insensitive to the fact that consumers in the Attr treatments already have information about the attribute x, and instead rate based on their own experience with the product.

Our experimental design separates out the rating decision from the sending decision so that we observe both the "pure" relationship between product quality and ratings independent of the cost of rating, as well as the relationship conditional upon raters paying that cost. The former is in some sense a more universal measure of how raters relate ratings to products, while the second more closely aligns with the ratings actually seen by buyers. Our final step is therefore to examine the sending decision of raters. Column 4 of Table 3 reports regression results on the decision to send a rating. In all treatments, raters are significantly more likely to send a rating for larger *ownvalue*. This aligns with our previous result that raters focus on communicating information about their own experiences.

6.2 Consumer Behavior

The second step in our analysis is to understand how consumers interpret ratings, and specifically how sensitive consumer willingness to pay is to observed ratings. Figure 2 shows the relationship between WTP and observed ratings in each treatment, while Table 4 reports regression results on consumer WTP decisions. Column 1 of Table 4 reports results across all treatments, where we see a positive effect of ratings on WTP, indicating that consumers do believe that higher rated products are more valuable. We

	(1)	(2)	(3)	(4)
	Choice of Rating (Tobit)			Choice to Send (LPM)
	All Ratings	Sent Ratings	Sent Ratings	All Ratings
Ownvalue	0.47***	0.44***		0.022***
	(0.030)	(0.067)		(0.0055)
Othervalue	0.069**	0.061		0.0071*
	(0.028)	(0.050)		(0.0040)
x			0.37***	
			(0.065)	
y			0.28***	
			(0.063)	
Pref	0.12	-1.01*		0.053
	(0.23)	(0.52)		(0.055)
Attr	0.24	0.56	0.90	0.028
	(0.23)	(0.52)	(0.66)	(0.053)
Pref × ownvalue	0.048	0.18**		0.016**
	(0.034)	(0.070)		(0.0078)
Pref × othervalue	-0.074**	-0.012		-0.0073
	(0.031)	(0.053)		(0.0052)
Attr × ownvalue	-0.021	0.0010		0.0063
	(0.034)	(0.072)		(0.0074)
Attr × othervalue	-0.0052	-0.076		-0.0066
	(0.031)	(0.052)		(0.0053)
Attr × x			-0.12	
			(0.087)	
Attr × y			-0.014	
			(0.095)	
Round	-0.014***	-0.024**	-0.019*	-0.0018*
	(0.0041)	(0.0010)	(0.011)	(0.0011)
Constant	0.46**	1.13**	0.53	0.058
	(0.22)	(0.52)	(0.54)	(0.039)
Observations	4040	1071	1071	4040

Standard errors in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 3: Rater decisions. Column 1 reports choice of rating for all ratings, Column 2 reports choice of rating for only sent ratings. Column 3 looks closely at the Attr treatments (x being common knowledge). Column 4 looks at the choice to send ratings. Column 1 - 3 use Tobit specifications. Column 4 uses a linear probability model. Bootstrapped standard errors based on 2000 replications.

	(1) All Ratings	(2) Pref	(3) Attr
Rating	0.93***	0.25***	0.59***
	(0.10)	(0.089)	(0.13)
Pref	0.95**		
	(0.43)		
Attr	0.75*		
	(0.44)		
Pref × rating	-0.26**		
	(0.12)		
Attr × rating	-0.13		
	(0.11)		
Sametype		-2.37***	
		(0.48)	
Sametype × rating		0.73***	
		(0.13)	
Ctypex			1.66***
			(0.57)
Ctypex × rating			-0.19
			(0.15)
x			0.39***
			(0.082)
x × rating			-0.012
			(0.016)
Round	0.032***	0.039***	0.029**
	(0.0077)	(0.011)	(0.011)
Constant	1.80***	4.20***	0.63
	(0.34)	(0.38)	(0.56)
Observations	4000	2000	2000

Standard errors in parentheses

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

Table 4: OLS results for consumer WTP, for consumers who observe ratings, and treating ratings as a continuous variable. Bootstrapped standard errors based on 2000 replications.

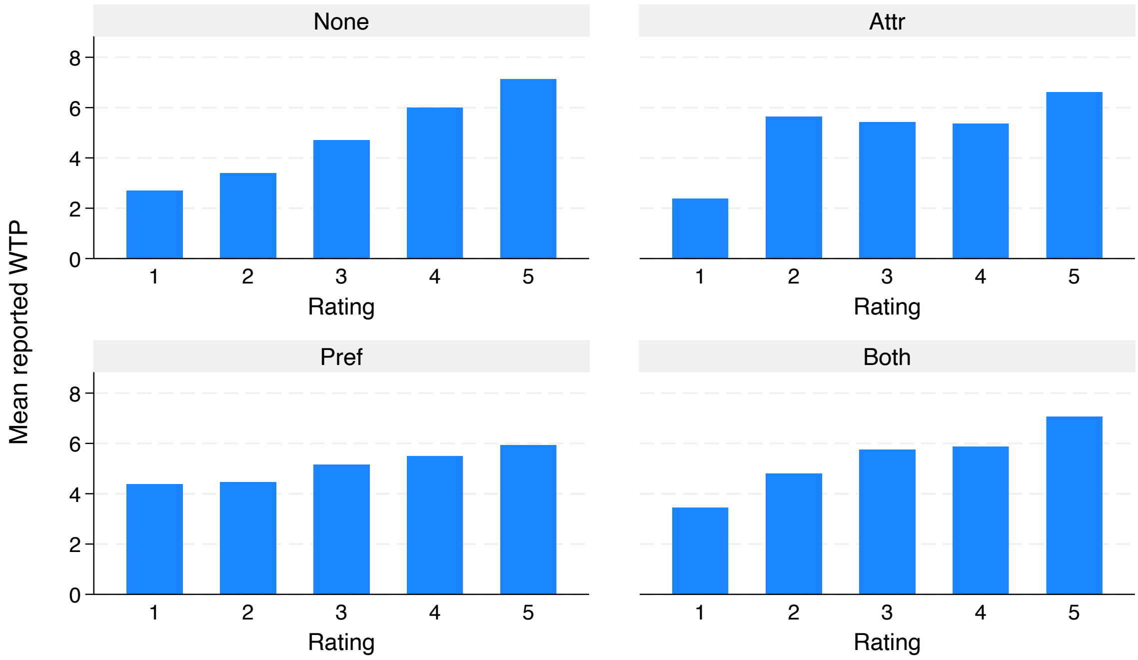

Figure 2: Mean WTP for consumers observing each rating, across treatments.

also find that ratings are less influential when rater preferences are public information, but there is no similar effect when x is common knowledge.

In order to test our specific predictions for the Pref and Attr treatments, we restrict columns 2 and 3 of Table 4 to only Pref and Attr treatments, respectively. The dummy variable *sametype* takes the value 1 if the consumer has the same preferences as their matched rater, and 0 if their preferences are different. Likewise, the variable *ctypex* takes the value 1 for X-type consumers and 0 for Y-type consumers.

Recall from Hypothesis 4 that in the Pref treatments, consumers who share the same preference as their rater should be more sensitive to the rating they observe. We test this in column 2 of Table 4 by interacting *sametype* with *rating*, and find that WTP is more sensitive to ratings when the consumer has the same preferences as the rater. Combined with our earlier result that raters rate based on their own preferences, we conclude both raters and consumers expect publicly known preferences to serve as a focal point for ratings. This serves as a partial answer to our research question, suggesting that knowledge of rater preferences can resolve the ambiguity resulting from heterogeneous preferences.

Our prediction in the Attr treatments is that consumers who prefer the known dimension, x, will be less sensitive to ratings than those who prefer the unknown attribute, y. We test this in column 3 of Table 4 by interacting *ctypex* with *rating*. Consumers who preferred x had a higher WTP on average, however we see no effect from interacting either consumer preferences or x with ratings, indicating there was no difference in terms of how consumers with different preferences interpreted ratings. This is contrary to our prediction, however it is consistent with our observed rater behavior. We conclude that both raters and consumers expect ratings to reflect the whole product, even when some attributes are publicly known.

We summarize our results as three main findings. First, raters tend to provide ratings based on their own tastes, contrary to our prediction that ratings are independent of rater preference. Second, when rater's preferences are public information, both raters and consumers interpret ratings as a reflection of rater preferences. Third, making some of a product's attributes publicly observable does not change the interpretation of ratings among raters or consumers, contrary to our predictions that ratings should only reflect unrevealed product quality. Overall, we find that rater and consumer interpretations of ratings are broadly consistent with one another in the presence of heterogeneous preferences.

7 Conclusion

Through the experience of others, ratings help consumers to learn about the quality of products prior to purchase. A rich existing literature has examined the generation and interpretation of ratings, typically focused on the motivation to rate products and the impact ratings have on consumer choices. Our results show how raters and consumers interpret ratings when they may disagree with one another as to what product characteristics are most important, and how that interpretation changes as additional information about rater preferences or product quality becomes available.

Regardless of other information available to consumers, raters tend to rate products based on their own preferences and consumers largely anticipate this behavior from raters. When consumers are informed of rater's preferences, we find ratings even more closely reflect those preferences. Consumers anticipate that raters will increase the focus on their own preferences, resulting in consumers becoming more sensitive to ratings if they share the same preferences as the rater, and less sensitive if their preferences differ. This suggests it may be beneficial to design rating systems which make rater preferences visible to consumers.

Providing consumers with partial information about a product's quality increases consumer willingness to pay, but does not change the meaning of ratings for raters or consumers. The increase in willingness to pay is not surprising, because consumers are mechanically better informed about products, regardless of any ratings. What is surprising is how access to this information does not change the meaning of ratings; although maximally informative ratings should avoid providing redundant information, raters do not change the way they select ratings based on whether product information is available to consumers. In other words, raters appear to continue rating the "whole product" rather than behaving strategically to maximize information transfer.

Our results provide insights into how heterogeneous preferences influence ratings, and how contextual information can affect the way ratings are interpreted. Given that raters tend to rate based on their own preferences and deprioritize the diverse preferences of others, our findings suggest that designers of rating systems should be mindful of the information consumers have about rater preferences. We also suggest that the information available in a ratings environment can change the way ratings are interpreted, as the information on each platform creates distinct contexts by which users interpret ratings. One possible implication is to caution against the comparability of ratings or reputations across different platforms, such as in Dellarocas et al. (2006).

This paper provides a first look at how ratings are interpreted in the presence of heterogeneous preferences, and there are many avenues to expand upon our findings. Our environment is intentionally a simple one, considering only products for which raters and consumers share ordinal preferences over each attribute, meaning that everyone can agree on what "better" means for each dimension of the product. Further research should examine other forms and contexts for heterogeneous preferences, such as how rating interpretations change when preferences are heterogeneous over both vertical and horizontal dimensions. Especially valuable would be an extension of our approach to studying written reviews, examining the degree to which raters focus on their own preferences when they have complete freedom to write arbitrarily detailed descriptions of products.

References

Becker, G. M., DeGroot, M. H., and Marschak, J. (1964). Measuring utility by a single-response sequential method. *Behavioral science*, 9(3):226–232.

Bolton, G., Greiner, B., and Ockenfels, A. (2013). Engineering trust: reciprocity in the production of reputation information. *Management science*, 59(2):265–285.

Bolton, G. E., Katok, E., and Ockenfels, A. (2004). How effective are electronic reputation mechanisms? an experimental investigation. *Management science*, 50(11):1587–1602.

Bowles, S. and Polania-Reyes, S. (2012). Economic incentives and social preferences: substitutes or complements? *Journal of Economic Literature*, 50(2):368–425.

Cabral, L. and Hortacsu, A. (2010). The dynamics of seller reputation: Evidence from ebay. *The Journal of Industrial Economics*, 58(1):54–78.

Cabral, L. M. (2000). Stretching firm and brand reputation. *RAND Journal of Economics*, pages 658–673.

Cai, H., Jin, G. Z., Liu, C., and Zhou, L.-a. (2014). Seller reputation: From word-of-mouth to centralized feedback. *International Journal of Industrial Organization*, 34:51–65.

Chakraborty, I., Kim, M., and Sudhir, K. (2022). Attribute sentiment scoring with online text reviews: Accounting for language structure and missing attributes. *Journal of Marketing Research*, 59(3):600–622.

Chen, D. L., Schonger, M., and Wickens, C. (2016). otree—an open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, 9:88–97.

Chen, Y., Harper, F. M., Konstan, J., and Li, S. X. (2010). Social comparisons and contributions to online communities: A field experiment on movielens. *American Economic Review*, 100(4):1358–98.

Chevalier, J. A. and Mayzlin, D. (2006). The effect of word of mouth on sales: Online book reviews. *Journal of marketing research*, 43(3):345–354.

Chiles, B. (2021). Shrouded prices and firm reputation: evidence from the us hotel industry. *Management Science*, 67(2):964–983.

Connors, L., Mudambi, S. M., and Schuff, D. (2011). Is it the review or the reviewer? a multi-method approach to determine the antecedents of online review helpfulness. In *2011 44th Hawaii International Conference on System Sciences*, pages 1–10. IEEE.

Crawford, V. P. and Sobel, J. (1982). Strategic information transmission. *Econometrica: Journal of the Econometric Society*, pages 1431–1451.

Dellarocas, C., Dini, F., and Spagnolo, G. (2006). *Designing reputation mechanisms*, page 446–482. Cambridge University Press.

Dellarocas, C., Zhang, X., and Awad, N. F. (2007). Exploring the value of online product reviews in forecasting sales: The case of motion pictures. *Journal of Interactive marketing*, 21(4):23–45.

Fradkin, A., Grewal, E., and Holtz, D. (2021). Reciprocity and unveiling in two-sided reputation systems: Evidence from an experiment on airbnb. *Marketing Science*, 40(6):1013–1029.

Gao, B., Li, X., Liu, S., and Fang, D. (2018). How power distance affects online hotel ratings: The positive moderating roles of hotel chain and reviewers' travel experience. *Tourism management*, 65:176–186.

Gesche, T. (2022). Reference-price shifts and customer antagonism: Evidence from reviews for online auctions. *Journal of Economics & Management Strategy*, 31(3):558–578.

Halliday, S. D. and Lafky, J. (2019). Reciprocity through ratings: An experimental study of bias in evaluations. *Journal of Behavioral and Experimental Economics*, 83:101480.

Healy, P. J. (2018). Explaining the bdm—or any random binary choice elicitation mechanism—to subjects. Technical report, mimeo.

Hoyer, B. and van Straaten, D. (2022). Anonymity and self-expression in online rating systems—an experimental analysis. *Journal of Behavioral and Experimental Economics*, 98:101869.

Hui, X., Saeedi, M., Spagnolo, G., and Tadelis, S. (2023). Raising the bar: Certification thresholds and market outcomes. *American Economic Journal: Microeconomics*, 15(2):599–626.

Jin, G. Z. and Kato, A. (2006). Price, quality, and reputation: Evidence from an online field experiment. *The RAND Journal of Economics*, 37(4):983–1005.

Lafky, J. (2014). Why do people rate? theory and evidence on online ratings. *Games and Economic Behavior*, 87:554–570.

Li, L., Tadelis, S., and Zhou, X. (2020). Buying reputation as a signal of quality: Evidence from an online marketplace. *The RAND Journal of Economics*, 51(4):965–988.

Luca, M. and Reshef, O. (2021). The effect of price on firm reputation. *Management Science*, 67(7):4408–4419.

Masterov, D. V., Mayer, U. F., and Tadelis, S. (2015). Canary in the e-commerce coal mine: Detecting and predicting poor experiences using buyer-to-seller messages. In *Proceedings of the Sixteenth ACM Conference on Economics and Computation*, pages 81–93.

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

Proserpio, D., Xu, W., and Zervas, G. (2018). You get what you give: theory and evidence of reciprocity in the sharing economy. *Quantitative Marketing and Economics*, 16(4):371–407.

Qiao, D., Lee, S.-Y., Whinston, A. B., and Wei, Q. (2020). Financial incentives dampen altruism in online prosocial contributions: A study of online reviews. *Information Systems Research*, 31(4):1361–1375.

Schneider, C., Weinmann, M., Mohr, P. N., and vom Brocke, J. (2021). When the stars shine too bright: The influence of multidimensional ratings on online consumer ratings. *Management Science*, 67(6):3871–3898.

Tadelis, S. (1999). What's in a name? reputation as a tradeable asset. *American Economic Review*, 89(3):548–563.

Woodside, A. G. and Davenport Jr, J. W. (1974). The effect of salesman similarity and expertise on consumer purchasing behavior. *Journal of Marketing Research*, 11(2):198–202.

Zhang, X., Luo, J., and Li, Q. (2012). Do different reputation systems provide consistent signals of seller quality: a canonical correlation investigation of chinese c2c marketplaces. *Electronic Markets*, 22:155–168.