American Journal of Business and Management
Vol. 3, No. 3, 2014, 161-167
DOI: 10.11634/216796061403546

Seeing Stars: How the Mere Presence of Ratings Influences Willingness to Pay

Steven Craig Huff, PhD

Utah Valley University, 800 W. University Pkwy – MS 288, Orem, UT 84058, USA
Email: huff@uvu.edu

This article investigates the effects of ordinal product ratings (i.e., product ratings such as stars, diamonds, etc.) when they are superfluous, meaning they are arbitrary and redundant. It finds that ratings do influence willingness to pay even when they are superfluous. When superfluous product ratings are included in a menu, they prompt individuals to categorize products by rating; this categorization exaggerates willingness to pay for products in the highest and lowest ratings tiers (i.e., at the extremes). In the study reported here, participants indicated their willingness to pay for multiple products in five product categories while the presence of superfluous ratings is manipulated. Results reveal an *expansion effect*; that is, the mere presence of superfluous product ratings in a menu can expand the range of willingness to pay for the products in the menu without influencing perceived quality. Results further reveal the natural consequence of the expansion effect, the *rating effect*; that is, changing a product's superfluous rating can change willingness to pay for that product, even when its quality remains constant. These findings suggest that prior research overstates the information effects of product ratings and that firms may be able to act more strategically when deciding: 1) whether to include ratings in their menus; 2) what decision rule they use to assign ratings; and 3) how to craft their product menus to maximize profits.

Keywords: product ratings, willingness to pay, product menus, judgment and decision-making

Introduction

RottenTomatoes.com (RT) reports the percentage of critics in the United States who write positive reviews of a movie. RT always posts a rating next to the percentage—a rating of ROTTEN if it is below 60% and a rating of FRESH otherwise. Presented alone, this rating could be helpful, but two of its characteristics reduce its value. First, the rating is arbitrary. This is evident from the website's FAQ which originally stated when the site was started, "Why 60%? We *feel* that 60% is a *comfortable minimum* for a movie to be recommended" (emphasis added; Dodson, 2013). Essentially, non-experts (the creators of RT are not movie critics) arbitrarily chose 60%. Second, the rating is superfluous because RT always presents it next to the percentage of positive reviews upon which it is solely based. Consequently, the rating adds no additional information about the quality of a movie.

Superfluous product ratings like the ones on RT are not common. Product ratings usually contain some information that a consumer cannot easily obtain otherwise. When Hotels.com posts a four-star rating for a Marriott hotel, it contains some information from their evaluation that is not conveyed in the online summary of hotel attributes and performance. The same is true for other expert ratings (e.g., J. D. Powers, Consumer Reports, MacWorld, etc.) and for user ratings (e.g., Amazon, Walmart, NewEgg, etc.). For example, suppose users rate two digital cameras at four stars each. Suppose further that they are similar in price and all attributes but resolution, and that one is a ten-megapixel camera and the other a twelve-megapixel camera. While the latter would show a higher megapixel count in the product description, the same rating on both cameras would suggest that users don't value the extra pixels. In this scenario, the ratings contain information about the cameras not contained in their product descriptions.

Since ratings usually provide some novel information, research investigating the influence of ratings on preference tends to attribute this influence to information effects without attempting to discover whether it also stems from additional sources. (Beaulieu, 2002; Jin & Sorenson, 2006; Scanlon et al., 2002; Wedig & Tai-Seale, 2002; Jin & Leslie, 2003). But some research suggests that even superfluous ratings—like those found on RT—may influence consumer preferences. Del Guercio & Tkac (2008) find that Morning Star ratings—mutual fund ratings on a discrete, five-star scale—are highly correlated with continuous performance measures

(i.e., Jensen's alphas, Sharpe measures, 3-year cumulative raw returns, etc.), but changes in morning star ratings correlate very little with these continuous measures. Using this, they show that a change in a fund's Morning Star rating influences investor flows significantly even when related continuous performance measures change very little. In a different paper, Figlio and Lucas (2004) employ an event study to investigate the response of house prices to school ratings in Gainsville, Florida. They cite Black (1999), who compares prices of adjacent houses that are divided by school zoning boundaries to demonstrate that school quality was already reflected in house prices before the rating system was introduced. They then show that when the school rating system was implemented, its influence was significant and substantial even though the ratings were based completely on test scores already available to the public.

While these studies suggest that superfluous ratings may influence preferences, they lack the experimental controls required to completely support this finding—a common problem with field experiments. Since Morning Star ratings can significantly lower search costs, it is plausible that many investors rely on the credibility of the Morning Star rating and do not seek the underlying information upon which it is based (including the continuous performance measures mentioned); consequently, the ratings can no longer be viewed as superfluous. A similar argument applies to house buyers in Florida. Once the school rating system was implemented, it is likely that many home buyers sought only school ratings without seeking the underlying test information upon which they were based.

This article closes the research gap by showing in an experimental setting with tight controls that superfluous ratings can and do influence consumer preferences. It further contributes to current literature by providing a theoretical explanation and supporting evidence for this influence.

One study is employed to demonstrate this phenomenon. In the study, participants reported their willingness to pay for four products in multiple hypothetical purchase situations. The results of the study demonstrate an *expansion effect;* that is, the mere presence of superfluous product ratings in a menu can expand the range of willingness to pay for the products in the menu. The study also demonstrates a *rating effect*; that is, changing a product's superfluous rating can change willingness to pay for that product, even when its quality remains constant.

This article continues as follows: first, it reviews current theory to develop hypotheses. Next, it describes the experiment used to test those hypotheses including the experimental design, data analysis, and results. Finally, it concludes with a discussion of marketing implications, limits to the research, and future research directions.

Theory and Hypotheses

Product quality (i.e., the sum of all product attributes) is often measured on a continuous scale, which means that there are no natural divisions in the product menu when evaluated solely on the basis of objective quality. In contrast, ratings (even when superfluous) provide a natural basis for dividing this continuous scale into discrete partitions, each containing only a subset of the menu. This categorization can enhance information processing efficiency as well as cognitive stability (Bruner, et al., 1956; Lingle, et al., 1984), but it also influences preferences in two important ways. The first consequence of is that these ratings categories prompt consumers to employ a phased-decision strategy (Bettman, 1979; Wright & Barbour, 1977) whereby in the first phase, they identify in the menu a subset of like-rated products on which to focus. In the second phase, they determine their willingness to pay by considering only these products in isolation (Kahneman & Miller, 1986; Leclerc, et al., 2005). Such a procedure for determining willingness to pay is comparable in the choice domain to the use of a consideration set when choosing from a large set of alternatives (Alba & Chattopadhyay, 1985; Hauser & Wernerfelt, 1990). This prompts the first hypothesis:

Hypothesis 1: Ratings matter: their mere presence (even when superfluous) prompts their use in determining willingness to pay.

The second important consequence of ratings-based categorization is that differences between products with different ratings and similarities between like-rated products both become exaggerated as the result of how categorical information is processed. Specifically, when the brain stores categorized information in memory, it focuses on similarities across objects within the same category and differences between objects of different categories (Cohen & Basu, 1987). This is cognitively more efficient than remembering all information about all objects. However, Cohen and Basu also show that the consequence of focusing on similarities and differences is that the perception of both similarities and differences becomes exaggerated. This exaggeration should bias willingness to pay for a product when it has a rating. Specifically, willingness to pay should increase as a product's rating increases, even when its objective quality

remains constant. Additionally, the inclusion of ratings in a menu should cause willingness-to-pay "gaps" between products of dissimilar ratings, resulting in an expansion of the willingness-to-pay range associated with the menu. This sets the stage for the second hypothesis:

Hypothesis 2: The ratings assignment rule matters: specifically, a higher (superfluous) rating for the same product will cause willingness to pay for that product to increase, despite that absolute quality remains unchanged.

Experimental Design and Hypothesis Testing

The experiment employs a 3 (rating assignment rule: CONTROL = No Ratings, UNIFORM ratings, and NON-UNIFORM ratings; between subjects) x 4 (menu position: 1, 3, 10, and 12) design. Each participant was asked to provide their willingness to pay for four products out of a menu of twelve products in five unique product categories presented in random order. The rating assignment rule varied between subjects and the target product in the menu varied within subjects.

For each product category, a purchase situation was described in detail, a product menu containing twelve products (in random order with respect to quality) was presented, an average price for the category was given, and willingness to pay was elicited for (when ordered by quality) the first, third, tenth, and twelfth products (hereafter referred to as Product 1, Product 3, Product 10 and Product 12, respectively). Each situation description emphasized that products in the menu were "essentially the same" except that their quality differed on one attribute (e.g., airlines differed by their average on-time arrival percentages). Each product menu contained objective quality levels on all twelve products for the differentiating attribute. Table 1 lists each of the product categories used in the survey and their corresponding differentiating quality attributes.

Table 1 – Product Categories

Product Category	Differentiating Attribute	Attribute Description
Airlines	On-time Record	% of on-time arrivals
Hotels	Complaints	% of customers who register official complaints
Cell Service	Coverage	% of local area covered
Auto Insurance	Settled Claims	% of claims settled satisfactorily
Laptops	Reliability	Probability of critical component failure in 5 years

To ensure the design would be robust to various product quality ranges, five quality distributions were used in the product menu. These were normally distributed, each with a mean of 70%. Each quality level across the five distributions was generated with the same z-scores[1] to ensure that the relative quality of each product remained constant in each distribution. Except in the CONTROL condition, each product in the menu was assigned one of the following four product ratings: four stars, three stars, two stars, or one star; subjects were truthfully informed that rating assignments were based solely on the differentiating attribute found in the product table for each category. This fact and the sentence, "Assume that these [products] are essentially the same except for the differences shown in the table below" in the situation description were included to emphasize the fact that the inclusion of ratings provided no novel information. In the UNIFORM condition, each rating was assigned to three products. In the NON-UNIFORM condition, the highest and lowest ratings were each assigned to two products, and the remaining two ratings were assigned to four products each. For each product in the menu, Table 2 shows for the product position, the product's quality z-score, its average quality across the five quality distributions used, and the rating assignments associated with each condition.

Table 2 – The Menu of Products and Rating Assignments

Product	Quality Z-Score	Mean Quality	Rating Assignments by Condition		
			Control	Uniform	Non-Uniform
1*	-1.831	82.9%	NA	4 Stars	4 Stars
2	-1.160	78.3%	NA	4 Stars	4 Stars
3*	-0.804	75.8%	NA	*4 Stars*	*3 Stars*
4	-0.550	73.8%	NA	3 Stars	3 Stars
5	-0.329	72.1%	NA	3 Stars	3 Stars
6	-0.109	70.6%	NA	3 Stars	3 Stars
7	0.090	69.2%	NA	2 Stars	2 Stars
8	0.306	67.7%	NA	2 Stars	2 Stars
9	0.539	66.1%	NA	2 Stars	2 Stars
10*	0.826	64.4%	NA	*1 Star*	*2 Stars*
11	1.181	61.9%	NA	1 Star	1 Star
12*	1.844	57.2%	NA	1 Star	1 Star

* willingness to pay was elicited for these products; italics used to highlight differences between the two treatment conditions.

Both treatment conditions assign each product a rating of one to four stars based on each product's quality, but because a different assignment rule is used in each condition, the design has the nice property that the rating for the third and tenth products in the menu changes across the two treatment conditions while the quality of these products across these conditions does not. This is in contrast to the first and last products in the menu where both the rating and quality remain constant across the two treatment conditions.

One-hundred and fifty-three undergraduates from a large university in the western United States were recruited from a large summer-session marketing class and paid a nominal fee to participate in the study. This provided 612 willingness to pay observations per product category for a combined total of 3,060 willingness to pay observations.

This experimental design employs a within-subjects factor (a product's position in the menu), which increases the amount of data that can be collected from subjects (when compared to a between subjects design). This advantage, however, comes with the cost of increased complexity in the econometric methods that must be used to analyze the resulting data. Specifically, the elicitation of four willingness to pay responses per subject in any given category violates the assumption that each observation is independent of other observations in the sample, an assumption that must be met in order to apply general linear models (i.e., OLS or ANOVA). For this reason, the experimental data are analyzed using a mixed-effects model with random intercept to control for the repeated measures just mentioned. Such a model is also called a hierarchical linear model or HLM (Raudenbush & Bryk, 2002) and can be thought of as the regression equivalent to repeated measures ANOVA.

Results and Discussion

Before analyzing the data with an HLM, it is interesting to see the raw results of the collected data. Table 3 compares the aggregated raw willingness to pay responses for Products 1, 3, 10, and 12 in the airline category. As shown, the willingness to pay responses in the control condition (when no ratings are included in the product menu) are compared to the combined responses of the two treatment conditions (when ratings are included in the product menu). Notice that the willingness to pay range in the "Without Ratings" column is significantly more compressed than the range in the "With Ratings" column. Willingness to pay data collected for the other four categories show similar patterns.

Table 3 – Airline WTP for All Products (1, 3, 10, and 12)

Product	Without Ratings	With Ratings
1	$420.39	$458.92
3	$385.20	$410.27
10	$333.37	$314.48
12	$297.35	$275.58

Table 4 compares the aggregated raw willingness to pay responses for only Products 3 and 10 in the airline category across the two treatment conditions that include product ratings (i.e., the two products which receive different product ratings in each treatment but have the same product quality across treatments). As shown, the willingness to pay responses in the control condition have been dropped and the figure compares only the two treatment conditions. Again, notice how willingness to pay for Product 3 in the UNIFORM condition is higher than in the NON-UNIFORM condition, and willingness to pay for Product 10 is lower. These differences correspond to a higher rating for Product 3 and a lower rating for Product 10 in the UNIFORM condition. Willingness to pay data collected for the other four categories show similar patterns.

Table 4 – Airline WTP for Products 3 and 10

Product	UNIFORM Condition	NON-UNIFORM Condition
3	$446.57	$373.88
10	$312.82	$316.14

Before presenting model estimates, it is important to confirm the appropriateness of the HLM over the general linear model. Likelihood-ratio tests were performed using the HLM as the unrestricted model and the analog general linear model as the restricted model for all five product categories for each HLM estimated. Every instance of this test is significant, suggesting that the HLM is indeed the more appropriate model for this analysis.

To test Hypothesis 1, Likelihood Ratio tests we performed that compared a basic HLM which included only product quality as the restricted model to the unrestricted HLM that included product quality as well as dummy variables (4 Stars, 3 Stars, 2 Stars, and 1 Star) to capture the possible effect of each product rating on willingness to pay. The LR tests yielded Chi-squared values (with 4 degrees of freedom) of 89.87, 87.59, 78.59, 53.17, and 84.68 for the airline, hotel, cell service, auto insurance, and laptop product categories, respectively. Each of these chi-square values indicates the LR test was highly significant at the 1% level, suggesting that the superfluous product ratings do influence willingness to pay. This result supports Hypothesis 1—suggesting that indeed the mere presence of product ratings, even when superfluous, influences willingness to pay. Table 5 shows the estimates from the unrestricted hierarchical linear model used to support Hypothesis 1. In the table, the intercept represents willingness to pay for Product 1 in the CONTROL condition which has no ratings. Consequently, significance for 4 Stars suggests that the presence of the 4 star rating significantly raises willingness to pay over the unrated case, despite that the rating is superfluous and product quality is constant across conditions.

Table 5 – Willingness to Pay Estimates (in Dollars)

Category	Airlines	Hotels	Cell Service	Auto Insurance	Laptops
Intercept	407.99**	203.42**	54.49**	952.83**	1,558.87**
Quality	3.77**	1.26**	0.68**	8.53**	18.56**
4 Stars	64.02**	35.60**	4.08*	55.51*	130.83**
3 Stars	14.62	-5.06	-1.35	5.75	-24.38
2 Stars	-19.40	-32.43**	-2.49	-35.69	-117.7**
1 Star	-49.56**	-47.41**	-6.12**	-74.34**	-212.62**

N=612; *, ** = 5%, 1% levels, respectively

Table 6 shows the estimates from the hierarchical linear model used to test Hypothesis 2. Here, the intercept represents mean willingness to pay for the 4-star rated product (Product 1). The significant coefficient for the 3-star rating suggests that the presence of the stars is influencing willingness to pay. In the Airline category, for example, giving a product a 3-star rating instead of a 4-star rating can drop willingness to pay by $55. Additional Likelihood ratio tests ($Chi^2(1)$ Airlines: 67.96***, Hotels: 59.11***, Cell Service: 39.50***, Auto Insurance: 33.70***, and Laptops: 64.24***) confirm that S2, and S1 are significantly different from each other, indicating that the presence of each

regressor in the HLM is warranted. The significant coefficients on S3, S2, and S1 support Hypothesis 2, namely, that an increase (decrease) in a product's rating leads to a significant increase (decrease) in willingness to pay, despite that quality remains constant.

Table 6 – Willingness to Pay Estimates (in Dollars)

Category	Airlines	Hotels	Cell Service	Auto Insurance	Laptops
Intercept	460.61**	233.88**	58.80**	1,008.04**	1,678.60**
Quality	2.05*	0.08	0.73**	8.47**	17.01**
3 Stars	-55.02**	-47.27**	-5.24**	-49.94**	-158.07**
2 Stars	-99.17**	-82.25**	-6.08**	-91.74**	-259.97**
1 Star	-134.10**	-102.34**	-9.55**	-130.35**	-358.67**

N=612; *, ** = 5%, 1% levels, respectively

Discussion

For the marketing practitioner, in addition to the existence of these effects, it is important to note the implications of these effects. Expressed as a percentage of the mean category prices, the addition of product ratings to the product menu increases willingness to pay for products with the highest rating by 8% to 16% and decreases willingness to pay for products with the lowest rating by 8% to 22% (i.e., the expansion effect). Similarly, as a percentage of mean category prices, using a different assignment rule to add an additional star to a product's rating (i.e. the rating effect) can lead to a 4% to 16% increase in willingness to pay for that product, despite that the quality of the product remains unchanged. These percentages are considerable and can have a significant impact on a firm's bottom line because they represent a change in revenue without a change in cost, thus their affect on gross margins and consequently net income are magnified. Manufacturers should consider these effects as they design, price, and choose distribution channels for their offerings; retailers must also consider these effects when choosing whether or not to include product ratings in their product menus.

Conclusion

This article investigates the direct effects of superfluous ratings in a product menu on willingness to pay for the products found in that menu. The results demonstrate 1) an *expansion effect*; that is, the presence of product ratings (though they provide no additional information) directly influences willingness to pay by expanding it. This focus on menu subsets leads to 2) a *rating effect*; that is, changing a product's superfluous rating can change willingness to pay for that product, even when its quality remains constant.

As with all research, these findings prompt the opportunity and need for further research. First, the experimental data were collected using hypothetical purchase situations. It would increase the validity of these findings if these studies could be run in real-world purchase situations where money is exchanged and products and services can be examined, sampled, purchased, and consumed. Second, this article capitalizes on an anomaly that occurs when information is converted from a ratio scale (e.g., objective quality information) to an ordinal scale (e.g., ratings). Other such questions of scale are interesting. For instance, ratings exist on an ordinal scale but are often treated as though they form an interval or ratio scale as is evidenced by reports of "average" star ratings on many on-line review sites such as Amazon.com or Dpreview.com. How does the star effect influence willingness to pay when star ratings are presented on an aggregate, continuous scale? Also, if ratings are viewed as an interval scale, how might changing the origin of the scale affect willingness to pay (i.e., willingness to pay for a products in a menu which contains products rated at one, two, or three stars compared to the same menu with products rated at three, four, and five stars)? Third, this work does not investigate an additional, possible explanation for the influence of relative quality on willingness to pay–the possibility that individuals have utility for status, meaning their position in society as signaled by their consumption behavior. In such a paradigm, it matters how much better one product is when compared to another because product superiority signals higher status. Further research is needed to investigate whether the possibility that consumers have utility for status could generate effects similar to those in this research.

Note

1. To generate the z-scores 6,000 draws from a normal distribution were sorted by their z-score and divided into twelve equal groups of 500 draws each. Within each group, these 500 z-scores were averaged to obtain one representative z-score for the group, resulting in twelve z-scores that were normally distributed in each product menu.

References

Alba, J. W., & Chattopadhyay, A. (1985). Effects of context and part-category cues on recall of competing brands. *Journal of Marketing Research*, 340-349.

Beaulieu, N. D. (2002). Quality information and consumer health plan choices. *Journal of Health Economics*, *21*(1), 43-63.

Bettman, J. R. (1979). *Information processing theory of consumer choice*. Addison-Wesley Pub. Co.

Black, S. E. (1999). Do better schools matter? Parental valuation of elementary education. *The Quarterly Journal of Economics*, *114*(2), 577-599.

Bruner, J. S., Goodnow, J. J., & George, A. (1956). Austin. A Study of Thinking. *New York: John Wiley & Sons, Inc*, *14*, 330.

Cohen, J. B., & Basu, K. (1987). Alternative models of categorization: toward a contingent processing framework. *Journal of Consumer Research*, *13*(4), 455.

Del Guercio, D., & Tkac, P. A. (2008). Star power: The effect of Morningstar ratings on mutual fund flow. *Journal of Financial and Quantitative Analysis*,*43*(4), 907.

Dodson, J. (2013) *Mind over Meta.* Retrieved from http://www.gamerevolution.com/features/mind_over_meta/2 on June 3, 2014.

Figlio, D. N., & Lucas, M. E. (2000). *What's in a grade? School report cards and house prices* (No. w8019). National Bureau of Economic Research.

Hauser, J. R., & Wernerfelt, B. (1990). An evaluation cost model of consideration sets. *Journal of consumer research*, *16*(4), 393.

Jin, G. Z., & Leslie, P. (2003). The effect of information on product quality: Evidence from restaurant hygiene grade cards. *The Quarterly Journal of Economics*, *118*(2), 409-451.

Jin, G. Z., & Sorensen, A. T. (2006). Information and consumer choice: the value of publicized health plan ratings. *Journal of Health Economics*, *25*(2), 248-275.

Kahneman, D., & Miller, D. T. (1986). Norm theory: Comparing reality to its alternatives. *Psychological review*, *93*(2), 136.

Leclerc, F., Hsee, C. K., & Nunes, J. C. (2005). Narrow focusing: Why the relative position of a good in its category matters more than it should. *Marketing Science*, *24*(2), 194-205.

Lingle, J. H., Altom, M. W., & Medin, D. L. (1984). Of cabbages and kings: Assessing the extendibility of natural object concept models to social things. *Handbook of social cognition*, *1*, 71-117.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (Vol. 1). Sage.

Scanlon, D. P., Chernew, M., McLaughlin, C., & Solon, G. (2002). The impact of health plan report cards on managed care enrollment. *Journal of health economics*, *21*(1), 19-41.

Wedig, G. J., & Tai-Seale, M. (2002). The effect of report cards on consumer choice in the health insurance market. *Journal of Health Economics*, *21*(6), 1031-1048.

Wright, P., & Barbour, F. (1977). *Phased decision strategies: Sequels to an initial screening*. Graduate School of Business, Stanford University.