The Effect of Rating System Design on Opinion Sharing

Completed Research Paper

Ying Liu
Arizona State University
300 E. Lemon St, Tempe, AZ 85287
Yingliu_is@asu.edu

Pei-yu Chen
Arizona State University
300 E. Lemon St, Tempe, AZ 85287
Peiyu.chen@asu.edu

Yili Hong
Arizona State University
300 E. Lemon St, Tempe, AZ 85287
hong@asu.edu

Yong Ge
University of Arizona
1130 E. Helen St, Tucson, AZ, 85721
yongge@email.arizona.edu

Abstract

Despite an intensive attention on extracting product features from textual reviews, there is limited understanding of the interplay between numerical ratings and textual reviews, and how rating system design moderates the interplay. In this study, we use data from two leading restaurant review websites with different rating systems to analyze the impact of rating system design on opinion sharing. We find that multi-dimensional (MD) ratings do not substitute textual reviews. Consumers tend to share longer opinions in textual reviews in a more objective way using the MD system. Results from randomized experiments corroborate that MD ratings do not substitute text reviews. Consumers' decisions to read reviews are not affected by the rating system. MD system primes consumers to generate a more comprehensive numerical overall rating of all dimensions as well as more comprehensive textual reviews. In addition, consumers are also found to use more neutral words in their textual reviews.

Keywords: Rating System Design, Opinion Sharing, DID

Introduction

Online review platforms allow consumers to share their opinions about products. Ubiquitous and accessible online reviews provide a wealth of information about goods and services to consumers in their search, evaluation, and choice of products. Literature has endorsed that both numerical ratings (Godes and Mayzlin 2004, Chintagunta et al. 2010, Rosario et al. 2016) and textual reviews (Archak et al. 2011, Ghose et al. 2012) have an impact on consumers' decision making, but they also play different roles. Hu et al 2014 suggests that consumers may use ratings to reduce the decision sets and use textual reviews to do further evaluation to arrive at a decision. Because numerical ratings require less cognitive effort and consumers resort to simplifying strategies and heuristics to arrive at a decision due to cognitive limitation. However, one potential issue of numerical ratings is they may not be representative of the information embedded in textual reviews. Ratings cannot comprehensively reflect information on different product attributes (Archak et al 2011), primarily because a product usually comprises of multiple attributes and consumers, who are heterogeneous (Li and Hitt 2008, Godes and Silva 2012), may form different levels of

preferences towards different product attributes. For example, a consumer may prefer high image quality than other attributes when he evaluates a camera. It is possible that he may provide a 5-star rating when the camera performs well in image quality, ignoring other attributes. On the other hand, he may elaborate his opinions on image quality in textual reviews or share his experience on all aspects of the camera. Although product attributes could be identified from textual reviews (Hu and Liu 2004, Archak et al 2011), there is limited understanding of the distribution of number of attributes covered in each review. It is also not clear how consumers choose to provide ratings and reviews, and the relationship between ratings and reviews: consumers may have taken into account all dimensions when giving a rating, or the rating may reflect only a particular dimension which matters the most to a consumer. Similarly, when providing textual reviews, consumers may focus on the dimension that drives the rating, or they may choose to provide additional information not necessarily reflected in their ratings; they may focus on only positive attributes or only negative attributes, or they may provide comprehensive reviews covering all dimensions. Given consumers use ratings and textual reviews differently and potentially at different stages in their decision making, ideally, it will be great to have comprehensive ratings and comprehensive reviews. The goal of this paper is to take a deeper look at how consumers choose to provide ratings and reviews, and how ratings and reviews are related, and most importantly, if multi-dimensional rating system (MD system) may help achieve the goal of having more comprehensive ratings and more comprehensive reviews. MD system allows a user to rate different dimensions/attributes of their product experiences. A multi-dimensional rating system is found to be more informative to users, reducing user uncertainty and leading to higher consumer satisfaction (Liu et al. 2014). Yet, it is not clear how the introduction of a multi-dimensional rating system affects the content of reviews. On one hand, reviewers may not find the need to write comprehensive and long reviews because they may think they already adequately expressed their opinions through the multi-dimensional ratings (substitution effect). On the other hand, reviewers may attempt to justify their ratings on different dimensions (justification effect), leading to a review that is comprehensive and covering all dimensions. Taken together, the reviews in a multi-dimensional rating system may become either longer or shorter in length, and either broader (cover more dimensions) or narrower (cover fewer dimensions) in terms of number of topics. In addition, the introduction of MD system may also affect linguistic features on each product dimension. Reviews could either be deeper (longer reviews) or more superficial (shorter reviews) on each product dimension. Consumers could also focus more on positive aspects or more on negative aspects. Given MD primes consumers of different aspects of their consumption experiences, it is also likely that MD reviews become more objective (or neutral). Bearing the above in mind, in this study, we are interested in answering the following questions:

RQ1: How ratings and reviews reflect consumers' heterogeneous preference?

RQ2: Do ratings complement or substitute textual reviews?

RQ3: How does rating system moderate the interplay?

We collected data on the same set of restaurants from *Yelp* and *TripAdvisor* and adopted the DID method to control for restaurant quality change. Our results suggest that MD ratings do not substitute text reviews. To the contrary, consumers tend to share more information in textual reviews in a more objective way using the MD system. MD reviews have greater breadth (more dimensions) and depth (longer). An experimental study is corroborated with the observational study to further uncover the mechanism. MD system primes consumers to generate a more comprehensive numerical overall rating of all dimensions. Our study makes a pioneering effort in establishing the value of rating system design on opinion sharing.

Theory and Hypothesis Development

While it is possible that consumers' preferences for product attributes be reflected in both numerical ratings and textual reviews in a single-dimensional rating system (SD system), several theories and prior findings suggest that numerical ratings may not fully reflect consumer experience on all product dimensions because consumers tend to place more weight on certain product attributes/dimensions toward which consumers have extreme feelings, either positive or negative. First, consumers are

motivated to share positive or negative WOM for impressions. Previous research (Chung and Darke 2006; Hennig-Thurau et al. 2004; Sundaram et al. 1998) find that people are more likely to share positive things because they want to be perceived as being positive. At the same time, people are also motivated to share negative things to show discriminating tastes because reviewers were seen as more intelligent, competent, and expert when they wrote negative as opposed to positive reviews (Amabile 1983). Second, consumers are motivated to share positive or negative WOM for emotion control. According to the Balance Theory (Heider 1946, 1958, Newcomb 1953), people have a basic desire for balance in their lives (Zajonc 1971). Thus, when experiencing a strong unbalance from either a strong positive or negative consumption experience, consumers may attempt to restore the equilibrium by expressing related positive emotions and negative feelings in reviews. This motive is referred as Homeostase Utility (Hennig-Thurau et al. 2004). For example, angry consumers (Wetzer et al. 2007) or dissatisfied customers (Anderson 1998) are more likely to share negative word of mouth to vent or to punish the company. Taken together, a consumer's overall satisfaction is likely to skew towards dimensions with the extreme sentiment, leading to ratings that are not comprehensive and are biased.

On the other hand, in an MD system, consumers may not only rate the restaurant overall but they also have an option to rate on different dimensions. MD system may cause consumers to report a more comprehensive overall rating. Compared to SD system, consumers are still motivated to share extreme feelings, however, their motivations of self-impression and emotion regulation could now be captured in dimensional ratings instead of the overall rating. And for the overall rating, the MD system may exert a priming effect (Neely 1977, Tipper 1985, Tulving and Schacter 1990). Consumers are primed with "multiple dimensions", which could remind consumers to take into account different dimensions, either positive or negative, and report a rating more representative of overall consumption experience. Therefore, we propose that:

H1: The overall ratings in MD system tend to reflect more dimensions of consumers' consumption experience compared to the overall ratings in SD system.

While a single numerical rating may not reflect consumers' overall experiences across multiple dimensions in an SD system, consumers could provide more information in textual reviews. Previous research has found that textual reviews contain information of different dimensions of product attributes using text mining approaches (e.g., natural language processing). Decker and Trusov (2010) considered rating heterogeneity and estimated the relative effect of product attributes and brand names on the overall evaluation of products. Ghose et al. (2012) estimated consumer demand and various product attributes using hotel reservation data and consumer-generated reviews and proposed a new ranking system that reflects the multidimensional preferences of consumers for products. Ghose et al. (2009) demonstrated that different dimensions indeed differentially affect the pricing power of sellers. However, these product attributes are extracted from the whole corpus of textual reviews. And it is still not clear how much information referred to product attributes is covered for each piece of textual review. Textual reviews do provide more details of product information, but it is possible that each piece of textual review only expand what consumers want to express in the numerical rating.

The introduction of MD system may lead to a change in the content generation in textual reviews in a few different ways. First, a substitution effect may exist. Given consumers' opinions have already been incorporated into numerical ratings on different dimensions, it is likely that consumers do not find the need to write a long review and elaborate on different dimensions. . In an SD system, since consumers can't express their opinions in one numerical rating, they may try to provide more details in their textual reviews, to make up for the deficiency in the single rating. For example, a consumer may feel bad about the service but good on other dimensions when he goes to a restaurant, and he may rate a 3-star in SD system to release bad emotions and explain in the textual review why he rates a 3-star and how he hates the service. And in MD system, he may just rate 3 on service and 5 on other dimensions and feel no need to explain in the textual review. Hence, we propose that:

H2a: Textual reviews substitute numerical ratings in MD system. Consumers tend to write shorter textual reviews in MD system than in SD system.

Alternative to the substitution effect proposed above, a justification effect could exist. MD system may lead to more content generation of textual reviews through the priming mechanism, since consumers are

primed with "multiple dimensions" in an MD system. In this case, we would expect consumers to write reviews which cover more dimensions in MD system compared to reviews in SD system. Much as dimensional ratings contain more information compared to a single rating in SD system, they also leave more information to be explained. Because consumers now provide both overall rating and dimensional ratings in MD system, it is possible that their overall ratings are not consistent with dimensional ratings. For example, in SD system, consumers only need to explain why they provide a 3-star overall rating. However, in MD system, consumers may attempt to explain why a 3-star on one attribute and a 4-star on the other attribute. According to the attribution theory, people tend to attach meaning to their behavior. In another word, they may tend to explain every dimensional rating they provide. Cognitive dissonance theory (Festinger 1957) suggests that people tend to seek consistency among their cognitions. When there is an inconsistency between attitudes or behaviors, something must change to eliminate the dissonance. When the inconsistency happens, consumers may feel like he needs to achieve consistency by rationalization and excuses, and he would explain in the textual reviews why he gives these dimensional ratings and the overall rating. And again, we would expect textual reviews in MD systems to be longer and cover more dimensions. Further, the number of product attributes listed itself could affect consumers' behavior. Sela and Berger (2012) argue that attribute numerosity is a heuristic cue for usefulness (Thompson et al. 2005), and according to the principle of multi-attribute diminishing sensitivity (Nowlis and Simonson 1996), increasing perceived usefulness through attribute numerosity should benefit more on hedonic than utilitarian options. That is, when choosing from different options, the number of attributes listed could imply more useful, and it benefits more on hedonic options. Hedonic options may be perceived more useful with more attributes listed. In our study, it is possible that consumers try to provide perceived useful information in textual reviews in SD system. And in MD system, the existence of dimensional ratings itself increases perceived usefulness which may lead to more information shared on attributes that are not "useful" in textual reviews. In this case, we would again expect more dimensions are covered in textual reviews in MD system than in SD system. Thus, we propose that:

H2b: Textual reviews complement numerical ratings in MD system. Consumers tend to write longer textual reviews in MD system than in SD system.

H3: On average, textual reviews in MD system are in greater breadth and depth than textual reviews in SD system.

Following the line of reasoning of cognitive dissonance, when people give a high overall rating, it is possible that he might not be highly satisfied with every dimension. For example, in MD system, consumers may rate the overall as 5, and 5-star for food, 4-star for service, 5-star for atmosphere and 5-star for value. In this case, the dimensional ratings are not consistent with the overall rating. And the consumer may try to rationalize his behavior and would explain on dimensions with lower ratings and we may expect an increase of negative emotions for high ratings.

H4: Negative emotions would be shared more in textual reviews in MD system than in SD system when overall rating is high.

Data

We address our research questions by studying restaurant ratings and reviews in different rating systems. We choose restaurants as our context because restaurants have well-known different dimensions of services (e.g., food and location) and attract significant attention in academic literature. We gathered data from two leading consumer review websites: Yelp.com (*Yelp*) and TripAdvisor.com (*TripAdvisor*). Like most review websites, *Yelp* provides a single-dimensional rating system on a scale of five stars. *TripAdvisor*, on the other hand, provides a multi-dimensional rating system, which allows not only overall ratings but also ratings for the dimensional characteristics of restaurants, such as food, service, and ambiance, using the same five-star rating scale. Figure 1 shows ratings of an identical restaurant, The Eddy in New York, on these two websites.

We used two customized web crawlers and collected data from these two websites. We obtained data for the identical restaurants of two review sites to eliminate restaurant differences and control for unobserved

quality changes in the restaurants. Therefore, the differences between the ratings in the two review systems for the identical restaurants cannot be attributed to the unobserved restaurant effect. We specifically match the restaurants according to restaurant names, addresses, and phone numbers in New York City. Finally, we obtained a sample of 698 restaurants. For each restaurant, we extracted the overall rating, dimensional ratings and reviews.

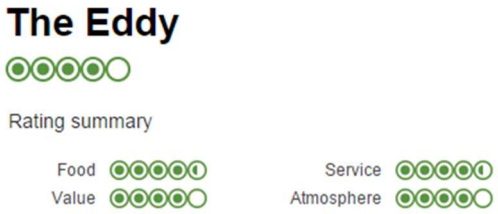
Yelp.com	
Tripadvisor.com	

Figure 1. Ratings of an Identical Restaurant on the Two Rating Websites

For each piece of textual review, we measured word count (WC), positive affect (PA) and negative affect (NA). We follow Golder and Macy (2011)'s approach to measure PA and NA by the proportion of positive emotion words, and the proportion of negative emotion words respectively. A higher value of PA denotes a higher portion of positive emotion words are used in the review. Besides, we move one step further to dig deeply what consumers are writing in text reviews. Specifically, we develop a novel machine learning method that models both textual reviews and overall ratings in a generative learning process. Different from the commonly used Latent Dirichlet Allocation (LDA) model (Blei et al.2003), our method takes both reviews and ratings as inputs, and the generation of review and rating influences each other through the overall generative process. Our method could automatically identify the topics embedded in all reviews and estimate the probability of each topic for reach review as well. These topic probabilities reflect the presence of topics in each review. As there are four predefined dimensions (i.e., food, service, value and atmosphere) in our restaurant review data, we further use some manually selected dimension-related words (e.g., salad and beef for food dimension) to align the automatically identified topics with these dimensions. In addition to the topic probability, our method could also generate a sentiment score on each dimension for very review. Such a score reflects the user's sentiment on each dimension expressed in his textual review. In the learning process of our method, the textual review and overall ratings jointly decide both topic probability and dimensional sentiment in a simultaneously way.

Consequently, we extract four topics which nicely correspond to the predefined four dimensions, food, service, atmosphere and value. The prevalence of each topic is represented by a dimension probability loading. For each piece of review, we take the log transformation of the product of word count and dimension probability loading to estimate the depth of each dimension. For example, if a review has 100 words, and the loadings of four dimensions are 0.25 respectively, then the depth of each dimension is log25. A higher number denotes higher depth that is more words are used to express opinions on this specific dimension. We use this measure instead of loading to control for word count. And then we compute the breadth of each review, basically, we use the number of dimensions mentioned in each review. Our method extracts four topics as well as their probabilities for each review. However, the probability loadings of some dimensions could be extremely low, we try to tease out these dimensions when measuring breadth of each review. We only count dimensions whose Z-scores of probability loadings are greater than -2, in another word, we don't consider those dimensions outside two standard deviations from the mean loading of this dimension across all reviews. To sum it up, a larger number suggests more dimensions are covered and a higher breadth of a review.

Research Setting and Methodology

Yelp adopts a single-dimension rating system, while *TripAdvisor* changed its rating system from single-dimension to multi-dimension in January 2009. To identify whether there is any effect of multi-dimensional rating system on emotion sharing, we compare variables of interest of *TripAdvisor* before and after the system change. However, there might be other reasons causing the change in variables of interest. For example, the quality of the restaurant might increase or decrease. In this case, we can't tell which factor cause the change. Here we take the difference in difference (DID) approach. We choose the exact same restaurants on *Yelp* as 'control group', therefore any trend on *Yelp* for each of these restaurants will serve as a proxy of change in restaurant quality. Besides, emotion sharing change at *TripAdvisor*, after controlling for the trend at *Yelp*, will be due to the change of the rating system.

We summarize this difference in difference approach below:

$$Y_{ij|k} = \beta_0 + \beta_1 * Time + \beta_2 * Time * Treat + \beta_3 * Treat + \beta_4 * \boldsymbol{X_{ijk}} + \alpha_i + \epsilon_{ijk}$$

where i indexes the restaurant, j indexes the position of the review in the review sequence for each restaurant and k denotes the website. Dependent variables are a list of variables including the overall rating, word count, positive and negative emotions, and breadth and depth from text mining. *Treat* is a dummy that equals one if the ratings are made on *TripAdvisor*, and zero if on *Yelp*. *Time* is a dummy that equals one if ratings are made after the system change, and zero if before the system change. The coefficient of the interaction term measures the difference caused by the change of the rating system, after controlling for changes in restaurant quality over time and systematic website differences. $\boldsymbol{X_{ijk}}$ is a vector of control variables. For example, we control ratings for word count and emotions. And we also control word count for emotions as emotions here are calculated as a portion of emotional words out of all words.

Results

Table 1. DID Analysis

	(1)	(2)	(3)	(4)
	Rating	WC	PA	NA
Time	-0.155***	-10.80***	0.644***	-0.0233
	(0.0147)	(1.316)	(0.0404)	(0.0146)
Treat	0.0837**	-99.41***	5.662***	0.158
	(0.0266)	(2.526)	(0.302)	(0.0822)
Time*treat	0.409***	58.48***	-6.031***	-0.154
	(0.0279)	(2.355)	(0.299)	(0.0804)
Rating		-11.25***	1.061***	-0.492***
		(0.290)	(0.0142)	(0.00816)
WC			-0.0169***	-0.000567***
			(0.000218)	(0.0000628)
N	143885	143885	143885	143885
Restaurant FE	Yes	Yes	Yes	Yes

Results are shown in Table 1. The dependent variable of the first column is the overall rating. And then in the following columns, dependent variables are word count, positive affect, and negative affect of each piece of textual review separately. The significant positive coefficient of the interaction term indicates that the change of the rating system from SD system and MD system significantly increase ratings by 0.409. The result is consistent with Liu et al (2014) where they find that the overall rating may increase due to increased information transfer efficiency when adopting MD system. And the results from the second column show that word count increases by almost 60 which indicates that consumers tend to write more when they use MD system. Our conjecture is that consumers tend to explain more on why they provide these dimensional ratings. They try to make their ratings more reasonable and credible. Results in column 3 and 4 show that positive affect decreases significantly, suggesting fewer positive words are being used,

while there is no significant change of negative affect. Overall, these results indicate that more neutral words are being used. The results are interesting because consumers are providing higher ratings which may suggest they are more satisfied, at the same time, consumers use fewer positive emotional words and more neutral words, suggesting that consumers are more objective. H2b is supported.

Table 2. DID Analysis of Breadth and Depth

	(1)	(2)	(3)	(4)	(5)
	Depth_food	Depth_service	Depth_value	Depth_atmo	Breadth
Time	-0.148***	-0.163***	-0.134***	-0.164***	-0.0529***
	(0.0129)	(0.0156)	(0.0150)	(0.0138)	(0.00354)
Treat	-1.896***	-1.816***	-1.653***	-1.361***	-0.938***
	(0.0434)	(0.0586)	(0.0418)	(0.0399)	(0.0325)
Time*treat	1.390***	1.788***	1.189***	1.197***	0.968***
	(0.0450)	(0.0535)	(0.0374)	(0.0410)	(0.0327)
Rating	-0.0791***	-0.154***	-0.133***	-0.00545	-0.0131***
	(0.00348)	(0.00389)	(0.00386)	(0.00352)	(0.00115)
Restaurant FE	Yes	Yes	Yes	Yes	Yes

Table 2 shows the results of how breadth and depth change after the adoption of MD system. The dependent variable of the columns (1) to (4) are the depth of four dimensions. A larger coefficient of the interaction term suggests consumers write more about this dimension. The significant positive coefficients of the interaction terms indicate that consumers tend to write more about all dimensions in MD system. MD system leads to greater depth of each dimension in text reviews. Consumers are not expanding their opinions on one or two specific dimensions, instead, they try to talk deeper in each dimension. The dependent variable of column (5) is the breadth of each review which is the number of dimensions covered in each review. The positive and significant coefficient of the interaction term suggests that on average MD reviews cover more dimensions. Results from table 2 suggest that MD reviews have greater breadth and depth than SD reviews. H3 is supported.

Rating Valence and Review

In this section, we want to analyze whether the effect is consistent across rating valence. Due to space constraint, we only report the coefficients of the interaction terms in table 3. Results from column (1) and (2) are consistent with table 1. Consumers tend to write more, and fewer positive words are used in all conditions, which suggest longer and more objective reviews in MD system. Results from column (3) show that fewer negative words are used when ratings are low. Again, the results suggest more objective reviews. Column (4) to (8) report consistent results with table 2 that review depth and breadth increase after the adoption of MD system across different ratings.

Table 3. DID Analysis_Rating Valence

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	WC	PA	NA	Depth_ food	Depth_ service	Depth_ value	Depth_ atmo	Breadth
Time*treat Rating<3	50.0*** (5.816)	-2.31*** (0.349)	-1.14** (0.435)	0.601*** (0.066)	0.96*** (0.086)	0.55*** (0.0617)	0.33*** (0.065)	0.517*** (0.0572)
Time*treat Rating=3	79.5*** (3.307)	-7.63*** (0.557)	0.013 (0.120)	1.219*** (0.0475)	1.596*** (0.048)	1.114*** (0.0476)	0.84*** (0.049)	1.073*** (0.0474)
Time*treat Rating>3	54.1*** (2.641)	-6.26*** (0.344)	-0.043 (0.0527)	0.89*** (0.038)	1.228*** (0.040)	0.70*** (0.030)	0.818*** (0.034)	0.916*** (0.0346)
Restaurant FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Robustness check

Within TripAdvisor Analysis

An alternative explanation to our findings may be that the differences are not from the system change but from website differences. That is, *Yelp* and *TripAdvisor* may attract a different set of audience who may have different writing styles. Therefore, in the following sections, we focus on within *TripAdvisor* analysis. And in case the user base of *TripAdvisor* itself may change due to the system change, we only focus on *TripAdvisor* data after the system change. Consumers of *TripAdvisor* are not forced to use MD system, instead, they could provide overall ratings with or without multi-dimensional ratings. This setting allows us to compare ratings and reviews from consumers who provide only SD ratings and who provide MD ratings.

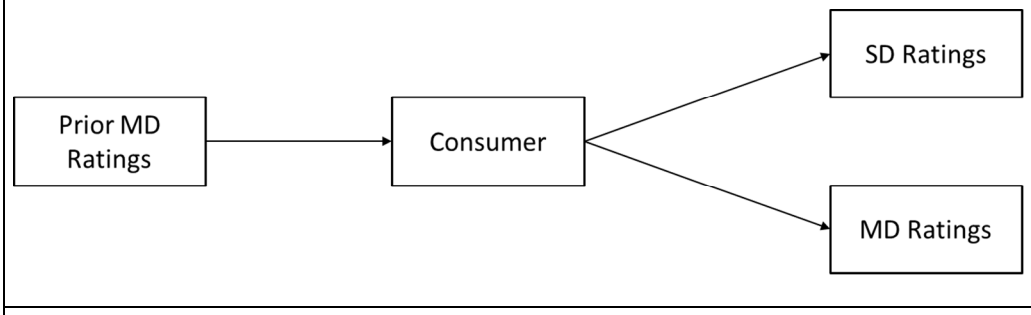


Figure 2. Within *Tripadvisor* After The System Change

We estimate the following equation.

$$Y_{ijr} = \beta_0 + \beta_1 * Multi + \beta_2 * X_{ijr} + \alpha_i + \gamma_r + \epsilon_{ijr}$$

Multi is a dummy variable that when it is set to one, it indicates when an overall rating is provided along with multi-dimensional ratings, and when it is zero, an overall rating is provided without multi-dimensional ratings. We also control for rating for word count and emotions. And we control for word count for emotions. And since we only use data within *TripAdvisor*, we are able to control both restaurant and reviewer fixed effect. The results shown in table 4 and table 5 are quite consistent with what we have in table 1 and 2. There is no difference of ratings within *TripAdvisor* as consumers obtained the same set of information according to Liu et al 2014. We could still see an increase in word count, a decrease in positive affect, no significant change in negative affect, and increase in both depth and breadth.

Table 4. Within *TripAdvisor* after the System Change

	Rating	WC	PA	NA
Multi	-0.02	32.5***	-0.6***	0.02
	(0.037)	(2)	(0.16)	(0.05)
Rating		-10.7***	1.05***	-0.4***
		(0.72)	(0.06)	(0.018)
WC			-0.02***	0.0001
			(0.001)	(0.0003)
Restaurant FE	Yes	Yes	Yes	Yes
Reviewer FE	Yes	Yes	Yes	Yes

Table 5. Within *Tripadvisor* after the System Change_Breadth and Depth					
	Depth_food	Depth_service	Depth_value	Depth_atmo	Breadth
Multi	0.503*** (0.0326)	0.580*** (0.0368)	0.360*** (0.0345)	0.358*** (0.0339)	0.0800*** (0.0105)
Rating	-0.0592*** (0.0117)	-0.160*** (0.0132)	-0.166*** (0.0123)	0.0718*** (0.0121)	-0.00410 (0.00377)
Restaurant FE	Yes	Yes	Yes	Yes	Yes
Reviewer FE	Yes	Yes	Yes	Yes	Yes

Table 6. Within *Tripadvisor* after the System Change_Rating Valence								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	WC	PA	NA	Depth_ food	Depth_ service	Depth_ value	Depth_ atmo	Breadth
Multi Rating=3	34.8* (16.07)	-0.898 (1.136)	-0.356 (0.518)	-0.0864 (0.278)	0.877** (0.293)	0.614* (0.295)	0.459 (0.334)	-0.0191 (0.0836)
Multi Rating=4	23.1*** (4.181)	-0.570 (0.383)	-0.130 (0.105)	0.44*** (0.0793)	0.496*** (0.0867)	0.28*** (0.0805)	0.34*** (0.0836)	0.08*** (0.0239)
Multi Rating=5	29.3*** (5.335)	-0.767 (0.528)	0.320** (0.123)	0.50*** (0.0952)	0.64*** (0.104)	0.39*** (0.103)	0.264* (0.105)	0.097** (0.0361)
Reviewer FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Restaurant FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

In table 6, we didn't report results when ratings are smaller than 3-star because of limited observations of ratings below three stars. From the results, we could tell that consumers tend to write more for medium to higher ratings. Fewer positive words are used in higher ratings. Reviews tend to cover more dimensions when the rating is high. And one interesting finding is that consumers are actually using more negative words in MD than in SD when the rating is 5 star.

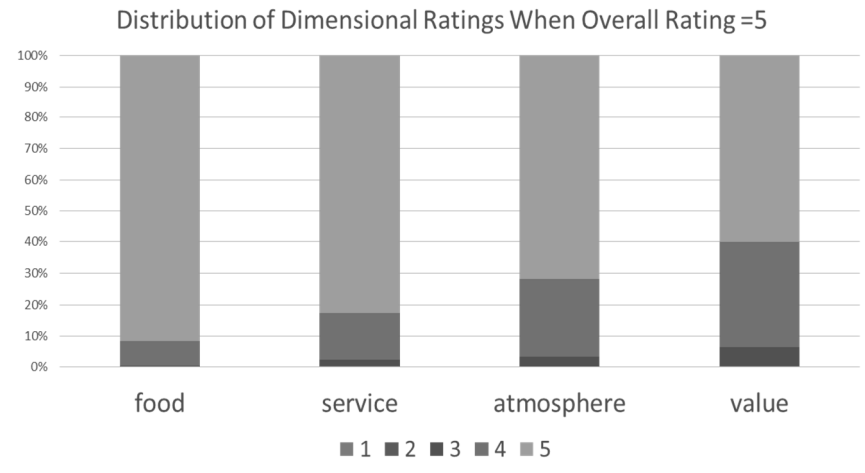


Figure 3. Distribution of Dimensional Ratings When Overall Rating = 5

We try to explain the result by looking through the distribution of dimensional ratings when the overall rating is 5-star in Figure 3. The first column shows that when the overall rating is 5, over 90% consumers also rate food as 5, and the fourth column shows that when the overall rating is 5, only about 60% consumers rate value as 5 and about 30% consumers rate value as 4 and about 10% consumers rate value as 3. The figure shows that even when consumers provide an overall rating of 5, they are still possible to rate different dimensions as less than 5, and they try to explain more about why they rate lower about these dimensions and therefore use more negative words compared to the case in SD system, given same 5-star rating. H4 is supported.

We are also able to analyze whether the discrepancies between the overall rating and the dimensional ratings lead reviewers to write longer reviews. The dependent variable is the word count and Discrepancy is the vector of the absolute differences between the overall rating and the dimensional ratings. We control for restaurant and reviewer fixed effects. The positive and significant coefficients of d_food and d_atmo suggest that people tend to explain more when there are discrepancies between food and the overall rating, and atmosphere and the overall rating.

$$WC_{ir} = \beta_0 + \beta_1 * Discrepancy + \alpha_i + \gamma_r + \epsilon_{ir}$$

Table 7. Review Length and Rating Discrepancy	
D_food	8.36*** (1.87)
D_service	1.85 (1.49)
D_atmo	5.94*** (1.43)
D_value	-0.96 (1.58)
Restaurant FE	Yes
Reviewer FE	Yes

Within Tripadvisor Consumers with both SD and MD

Some may also argue that self-selection issue may exist. Consumers self-select to use either SD or MD. It is possible that consumers who tend to write longer reviews would tend to use MD ratings. In this section, we not only focus on within *Tripadvisor* data but also only consider reviewers who provide both SD and MD reviews. We didn't observe any time trend that consumers would use SD first and then stick to MD. That is, empirically, users just switch between MD and SD randomly. Results are shown in Table 8 to 10. Note that in table 10, we report results when ratings are greater than 3 due to data limitation. Again, we see similar results as in previous sections.

Table 8. Consumers with Both SD and MD Within *TripAdvisor* after the System Change				
	Rating	WC	PA	NA
Multi	-0.00776 (0.0392)	31.63*** (2.168)	-0.652*** (0.180)	0.0309 (0.0574)
Rating		-10.75*** (1.153)	1.011*** (0.0935)	-0.382*** (0.0297)
WC			-0.0216*** (0.00166)	0.000108 (0.000527)
Restaurant FE	Yes	Yes	Yes	Yes
Reviewer FE	Yes	Yes	Yes	Yes

Table 9. Consumers with Both SD and MD Within *TripAdvisor* after the System Change_Breadth and Depth					
	Depth_food	Depth_service	Depth_value	Depth_atmo	Breadth
Multi	0.507*** (0.0370)	0.566*** (0.0414)	0.376*** (0.0381)	0.335*** (0.0378)	0.0774*** (0.0121)
Rating	-0.0601** (0.0197)	-0.190*** (0.0220)	-0.152*** (0.0202)	0.0913*** (0.0201)	-0.00569 (0.00645)
Restaurant FE	Yes	Yes	Yes	Yes	Yes
Reviewer FE	Yes	Yes	Yes	Yes	Yes

Table 10. Consumers with Both SD and MD Within *TripAdvisor* after the System Change _Rating Valence								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	WC	PA	NA	Depth_ food	Depth_ service	Depth_ value	Depth_ atmo	Breadth
Multi	21.38***	-0.566	-0.184	0.44***	0.467***	0.233*	0.251*	0.0623*
Rating=4	(5.061)	(0.488)	(0.124)	(0.0997)	(0.107)	(0.102)	(0.105)	(0.0305)
Multi	29.25***	-1.175	0.652***	0.48***	0.577***	0.253	0.463**	0.101
Rating=5	(8.231)	(0.701)	(0.168)	(0.142)	(0.148)	(0.146)	(0.158)	(0.0590)
Reviewer FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Restaurant FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Experiment

We also conduct experiments to help understand how consumers refer their preferences into ratings in SD and MD. We try to mimic an environment where consumers have the same consumption experience while using different rating systems. Specifically, subjects were asked to read someone else's dining experiences and then rate the restaurant using the two rating systems. Subjects were first shown four pieces of reviews. We choose a relatively small number to reduce cognitive burdens and make sure that all reviews have been read and all subjects obtain the same set of information. Subjects were told that all these reviews are authentic and they need to read all and consider these reviews as their own consumption experiences. And then subjects were randomly divided into four groups, SD, SD with priming, MD with overall rating first, and MD with dimensional rating first. Subjects in the first group were asked to rate the restaurant using SD system. Subjects in the second group were asked to think about different attributes of the restaurant (for example, service, food, ambiance, etc) and then rate the restaurant using SD system. Subjects in the third group were asked to first rate the restaurant on the overall rating, and then rate different dimensions of the restaurant (food, service, atmosphere, and value). Subjects in the fourth group were asked to first rate different dimensions of the restaurant (food, service, atmosphere, and value), and then rate the restaurant on the overall rating. Time used to read and rate reviews are recorded and subjects who didn't spend enough time will be teased out of the sample. Besides, on the same page, subjects will be asked to indicate the month of the day when they take the experiment, and those who didn't provide the correct answer will be teased out. And then, all subjects were asked to indicate what attributes (food, service, atmosphere, value, and other) have been considered when generating the overall rating and the importance of each of these attributes by allocating 100 points among the attributes. In total, we receive 2682 responses passing manipulation test out of 2745 responses. Results are shown in Figure 5 using 45 seconds as the cutoff which leads to 1615 valid responses. And we use different cutoffs of time spent on reading and rating reviews and the results are consistent.

Figure 4 shows the average percentages of different dimensions used to generate the overall rating. The results suggest that consumers do put different weight on different dimensions. The results also suggest a higher diversity of weight of different dimensions in SD systems than in MD systems. For example, consumers put more weight on food in SD compared to MD, but less weight on atmosphere and value.

MD system primes consumers to generate a more comprehensive evaluation of all dimensions. H1 is supported.

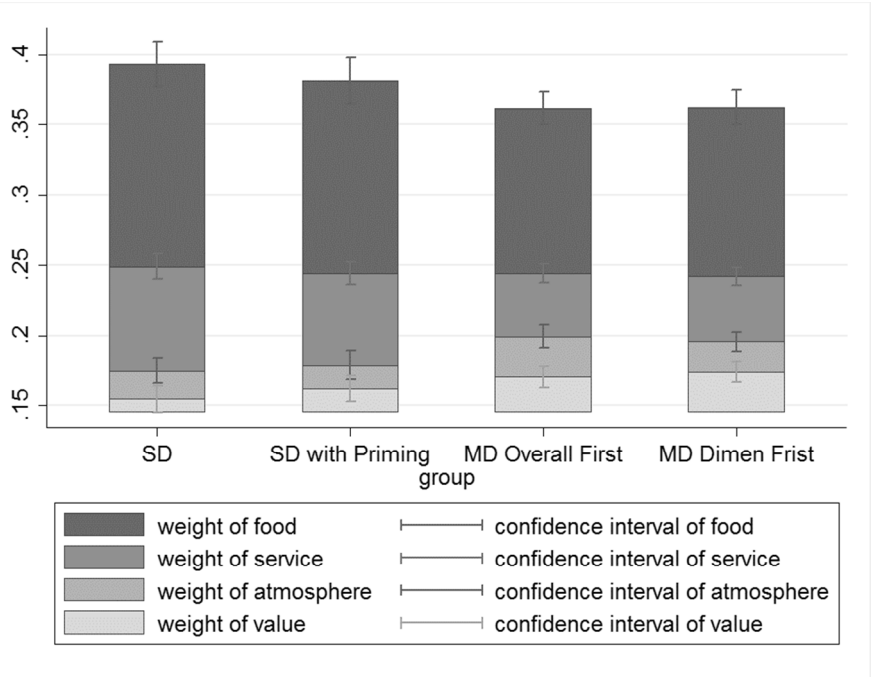


Figure 4. The Priming Effect on Dimensional Ratings

Another interesting and related question is whether the use of MD system reduces consumers' likelihood to read text reviews. We conduct another experiment in which respondents were traced if they clicked to read reviews of the restaurant after the rating information was provided to them. Subjects were first primed about a scenario that they will go for lunch near campus. Subjects were then shown four restaurants, two of which came with SD ratings (Restaurant 1 and 2), while the other two with MD ratings (Restaurant 3 and 4). Restaurant 1 and 2 were provided with only an overall rating while Restaurant 3 and 4 were presented with MD ratings. The overall ratings of Restaurant 2 and 4 are higher than those of Restaurant 1 and 3. Besides, for each restaurant, respondents were asked whether they want more information about the restaurant. If respondents answered yes, information concerning price level, restaurant description would be provided, and a further question of whether they want to read more text reviews would be asked. And the text reviews were shown in random order if respondents chose to read the text reviews. The display of information is similar to what one would see on the website. The respondents were then asked if they would choose to have lunch at each of the four restaurants. After the choice had been made, respondents were asked to answer a list of questions related to demographics, etc. The results show that on average, less than 30% of participants chose to read text reviews. The results also show that consumers' decisions to read reviews are not affected by the rating system. That is, the likelihood to read reviews is comparable in SD and MD systems. These results provide support that MD ratings do not substitute text reviews.

Discussion

We corroborated an observational study with an experimental study to examine how consumers reflect their overall consumption experience in ratings and reviews in different rating systems. Our results suggest that MD ratings do not substitute text reviews. Consumers in an MD system tend to share more information and cover more dimensions in textual reviews in a more objective way. A natural question following is that are higher depth and breadth reviews really helpful? Consumers read textual reviews

rather than relying simply on summary statistics (Chevalier and Mayzlin 2006) to resolve their uncertainty about product attributes (Pavlou and Dimoka 2006). Review depth has a positive effect on the helpfulness of the review (Mudambi and Schuff 2010). There is limited understanding of how review breadth impact review helpfulness which could be a potentially interesting topic for a future study. Results from randomized experiments corroborate that MD ratings do not substitute text reviews. Consumers' decisions to read reviews are not affected by the rating system. MD system primes consumers to generate a more comprehensive numerical overall rating of all dimensions as well as more comprehensive textual reviews. In addition, consumers are also found to use more neutral words in their textual reviews. Future research will dig further impact on other linguistic features and robustness checks of within-reviewer and between-reviewer variation. Our study contributes to rating system design and provides a better understanding of how ratings and reviews reflect consumers' experiences, and our findings also increase online retailers' understanding of the role rating system play in opinion sharing.

Reference

Benbasat, I., and Zmud, R. W. 2003. "The Identity Crisis within the IS Discipline: Defining and Communicating the Discipline's Core Properties," *MIS Quarterly* (27:2), pp. 183-194.

Amabile, T.M., 1983. Brilliant but cruel: Perceptions of negative evaluators. Journal of Experimental Social Psychology, 19(2), pp.146-156.

Archak, N., Ghose, A. and Ipeirotis, P.G., 2011. Deriving the pricing power of product features by mining consumer reviews. Management Science, 57(8), pp.1485-1509.

Babić Rosario, A., Sotgiu, F., De Valck, K. and Bijmolt, T.H., 2016. The effect of electronic word of mouth on sales: A meta-analytic review of platform, product, and metric factors. Journal of Marketing Research, 53(3), pp.297-318.

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

Chintagunta, P.K., Gopinath, S. and Venkataraman, S., 2010. The effects of online user reviews on movie box office performance: Accounting for sequential rollout and aggregation across local markets. Marketing Science, 29(5), pp.944-957.

Chung, C.M. and Darke, P.R., 2006. The consumer as advocate: Self-relevance, culture, and word-of-mouth. Marketing Letters, 17(4), pp.269-279.

Decker, R. and Trusov, M., 2010. Estimating aggregate consumer preferences from online product reviews. International Journal of Research in Marketing, 27(4), pp.293-307.

Festinger L A .1957. Theory of Cognitive Dissonance, Stanford University Press, Stanford, CA

Golder, S.A. and Macy, M.W., 2011. Diurnal and seasonal mood vary with work, sleep, and day length across diverse cultures. Science, 333(6051), pp.1878-1881.

Heider, F., 1946. Attitudes and cognitive organization. The Journal of psychology, 21(1), pp.107-112.

Heider, F., 1958. The psychology of interpersonal relations, New York: Wiley

Ghose, A. and Ipeirotis, P.G., 2011. Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics. IEEE Transactions on Knowledge and Data Engineering, 23(10), pp.1498-1512.

Ghose, A., Ipeirotis, P.G. and Li, B., 2012. Designing ranking systems for hotels on travel search engines by mining user-generated and crowdsourced content. Marketing Science, 31(3), pp.493-520.

Godes, D. and Mayzlin, D., 2004. Using online conversations to study word-of-mouth communication. Marketing science, 23(4), pp.545-560.

Godes, D. and Silva, J.C., 2012. Sequential and temporal dynamics of online opinion. Marketing Science, 31(3), pp.448-473.

Newcomb, T.M., 1953. An approach to the study of communicative acts. Psychological review, 60(6), p.393.

Hennig-Thurau, T., Gwinner, K.P., Walsh, G. and Gremler, D.D., 2004. Electronic word-of-mouth via consumer-opinion platforms: what motivates consumers to articulate themselves on the internet?. Journal of interactive marketing, 18(1), pp.38-52.

Hu, N., Koh, N.S. and Reddy, S.K., 2014. Ratings lead you to the product, reviews help you clinch it? The mediating role of online review sentiments on product sales. Decision support systems, 57, pp.42-53.

Hu, M. and Liu, B., 2004, August. Mining and summarizing customer reviews. In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 168-177). ACM.
Li, X. and Hitt, L.M., 2008. Self-selection and information role of online product reviews. Information Systems Research, 19(4), pp.456-474.
Liu, Y., Chen,P.Y. and Hong, Y., 2014. The Value of Multi-dimensional Online Rating Systems: An Information Transfer View. Proceedings of the 35th International Conference on Information Systems (ICIS), Auckland, New Zealand.
Mudambi, S.M. and Schuff, D., 2010. What makes a helpful online review? a study of customer reviews on amazon. com. MIS Quarterly, 34(1), pp.185-200.
Neely, J. H. 1977. "Semantic priming and retrieval from lexical memory: Roles of inhibitionless spreading activation and limited-capacity attention," Journal of experimental psychology: general (106:3), pp. 226.
Nowlis, S.M. and Simonson, I., 1996. The effect of new product features on brand choice. Journal of marketing research, pp.36-46.
Packard, G.M. and Wooten, D.B., 2013. Compensatory knowledge signaling in consumer word-of-mouth.
Pavlou, P.A. and Dimoka, A., 2006. The nature and role of feedback text comments in online marketplaces: Implications for trust building, price premiums, and seller differentiation. Information Systems Research, 17(4), pp.392-414.
Sela, A. and Berger, J., 2012. How attribute quantity influences option choice. Journal of Marketing Research, 49(6), pp.942-953.
Sundaram, D.S., Mitra, K. and Webster, C., 1998. Word-of-mouth communications: A motivational analysis. NA-Advances in Consumer Research Volume 25.
Tipper, S. P. 1985. "The negative priming effect: Inhibitory priming by ignored objects," The Quarterly Journal of Experimental Psychology (37:4), pp. 571-590.
Thompson, D.V., Hamilton, R.W. and Rust, R.T., 2005. Feature fatigue: When product capabilities become too much of a good thing. Journal of marketing research, 42(4), pp.431-442.
Tulving, E. and Schacter, D.L., 1990. Priming and human memory systems. Science, 247(4940), pp.301-306.
Wetzer, I.M., Zeelenberg, M. and Pieters, R., 2007. "Never eat in that restaurant, I did!": Exploring why people engage in negative word-of-mouth communication. Psychology & Marketing, 24(8), pp.661-680.
Zajonc, R.B., 1960. The concepts of balance, congruity, and dissonance. Public Opinion Quarterly, 24(2), pp.280-296.