THE RELIABILITY OF ONLINE REVIEW HELPFULNESS

Yun Wan
School of Arts & Science
University of Houston – Victoria, United States
wany@uhv.edu

Makoto Nakayama
College of Computing and Digital Media
DePaul University, United States
MNakayama@cdm.depaul.edu

ABSTRACT

Many online reviews have a helpfulness rating, and such ratings are being widely used by online shoppers for shopping research. Researchers also use them as a review quality benchmark. However, there is scant research about the reliability of such ratings. This paper explores the reliability of helpfulness ratings and their resistance to manipulations. We found that the existing helpfulness ratings for most helpful reviews are inflated and significantly higher than ratings we collected from a random population due to online shopper self-selection behavior. We also found existing helpfulness ratings for most helpful favorable reviews have an anchoring effect on subsequent votes, thus could be potentially manipulated to boost sales. In contrast, ratings for most helpful critical reviews have a counter-anchoring effect due to risk aversion, thus could backfire if manipulated. Implications and future research are discussed.

Keywords: Online review; Helpfulness; Amazon.com; User generated content; B2C ecommerce

1. Introduction

User-generated online reviews (online reviews hereafter) are becoming an essential component of B2C ecommerce. Online reviews mainly serve two functions in electronic commerce. One is to help online shoppers evaluate products and services before making purchase decisions [Park, Lee et al. 2007]. The other is informant [Clemons, Gao et al. 2006], which allows consumers to become familiar with a product or service even though they do not have an immediate intent to purchase [Chen and Xie 2008]. Online reviews can offer important value to customers [Mudambi and Schuff 2010]. Empirical studies done in the past ten years report that popular reviews have strong influences not only on commodities [Zhu and Zhang 2010] and new products [Cui, Lui et al. 2012] but also on services [Ye, Law et al. 2011].

Amazon.com revolutionized many online review features to enhance consumers' shopping experiences. For example, once a new online review has been posted and read by a registered shopper, the shopper can vote on its helpfulness by simply clicking the *Yes* or *No* button under the review content. The aggregated number of *Yes* vote and total votes a review received are then updated and displayed at the top of the review content as an indicator of helpfulness.

Based on the aggregated helpfulness votes a review receives, amazon.com could use its proprietary computer algorithms to automatically rank and sort out those *most helpful* reviews and feature them at the top of the review section. This simple feature seems very helpful when the number of reviews keeps increasing and consumers feel difficult to go through even a small percentage of them. Shoppers could then spend their limited product-research time on the most helpful ones to avoid information overload [Maes 1994]. Gradually, because of the market share and influence of amazon.com, this voting-for-helpfulness feature was not only being adopted by many other online retailers, but also being utilized by many researchers as a *de facto* review quality standard [Mudambi and Schuff 2010, Ghose and Ipeirotis 2011, Korfiatis, García-Bariocanal et al. 2012].

Helpfulness votes are important to help consumers on product research and making purchase decisions. It also serves as a benchmark for academic studies on online review. So understanding their reliability and resistance to manipulation could help us better utilize this feature. Several studies already identified bias in online reviews [Li and Hitt 2008, Kapoor and Piramuthu 2009, Cui, Lui et al. 2012, Purnawirawan, Dens et al. 2012]. The voting for helpfulness of reviews may also suffer from similar or related biases. We need to explore the impact of such bias on the helpfulness rating outcomes.

In addition to bias, there are increasing concerns about the review manipulation by interested parties like product manufacturers. Ordinary consumers may not be aware of online review manipulation but such practices were observed by both small businesses owners, like those listed on Yelp.com [Banks 2013], and researchers as reflected in recent studies [Mukherjee, Liu et al. 2012]. In September 2013, a few online review-rigging firms have been persecuted by the Court of New York State [Chappell 2013], which became the first law punishment on such practice. Since online review manipulations have strong financial incentive and were being systematically conducted, it is reasonable to suspect those firms to boost or bust certain reviews or to manipulate the helpfulness vote, which makes the helpfulness ratings unreliable. On the other hand, voting for helpfulness is a crowdsourcing continuous process. An online review-rigging firm may be able to manipulate the helpfulness votes for a product temporarily but it may not be economically viable to do so in the long term, especially for those products receiving votes continuously. So a helpfulness rating may self-correct such manipulation gradually.

In whichever case, the most helpful reviews featured for popular products on amazon.com are the most influential reviews on consumers. They are also likely the primary targets of manipulation, because they are not only the first review, but also, probably, the only review being used by many online shoppers to make product research, hence most influential on sales. Since these most helpful reviews are being identified and ranked by helpfulness ratings they receive, it is important to explore the reliability of such ratings, including whether there is bias in its helpfulness rating and how good is their resistance to short-term manipulations.

For the remainder of the paper, we first review existing literature and present our hypotheses, then explain the two experiments we conducted to explore above research questions, and finally discuss our findings and their implications.

1.1 Literature Review and Hypotheses

Though the abundance of online reviews provides convenience to online shoppers, they also pose a challenge: If there are many reviews for a product, they could overload online shoppers cognitively before the online shopper finds a truly helpful one [Wan, Menon et al. 2007]. Thus, online retailers have to help online shoppers to filter those reviews and identify quality reviews effectively. Obviously, established online portals like amazon.com were among the first group of online retailers facing such challenges. As a result, amazon.com gradually introduced a helpfulness voting filtering mechanism. This allows online shoppers not only to write a review for a product, but also to allow others to rate the helpfulness of reviews. Those reviews receiving the highest ratio of *Yes* votes against overall votes (its helpfulness rating) are featured on product page. Amazon.com further differentiates the *most helpful* reviews into two categories: the *most favorable* and the *most critical* reviews, depending on the product rating given by a reviewer. The most favorable review gives the product 4 to 5 stars while the most critical review gives the review 1 to 3 stars. The helpfulness voting and aggregated rating play an important role in consumer online shopping. Next we review existing literatures on how consumers make voting decisions on review helpfulness.

1.2 The Potentially Inflated Review Helpfulness

The normative decision-making model assumes consumers vote on the helpfulness of a review independently - that is they cast their votes based on review quality formed in their own short and long-term memory only. Even under this assumption, the voting could be influenced by other reviews, because though a memory that is used for voting could be an impression a consumer has of the overall helpfulness of the review after reading it (short term memory), it could be the helpfulness compared with other reviews read by the shopper in the same or previous period of time (short and long term memory). When we include the influence of other reviews in the decision factors, a voting process becomes complexity phenomenon and susceptible to orders of reviews published. The dynamics of such a process has been systematically studied in several existing literatures, including the online product reviews themselves [Kapoor and Piramuthu 2009], as well as on an artificial music market experiment about how hit songs became hits [Salganik, Dodds et al. 2006]. As a result, its eventual outcome - in this case, the most helpful review with the highest rating - is path-dependent and susceptible to initial conditions and a wide range of environmental factors [Brian Arthur, Ermoliev et al. 1987].

From an individual consumer’s perspective, both the review itself and the specific voters could influence his or her voting on the helpfulness of a review.

Ghose and Ipeirotis [2011] found that the extent of subjectivity, informativeness, readability, and linguistic correctness in reviews could influence perceived usefulness. Reviewer attributes also have impact. Connors et al. [2011] found that reviews written by a self-described expert are perceived as more helpful than those that are not. The impacts of these factors are different in extent. Review readability had a greater effect on the helpfulness ratio of a review than its length [Korfiatis, García-Bariocanal et al. 2012]. We also need to consider the interaction effect of these factors. For example, though both review valence and length have positive effects on review helpfulness, the product type (i.e., experiential vs. utilitarian product] moderates these effects [Pan and Zhang 2011]. Sometimes, such effect could be subtle. For example, longer reviews are not necessarily always better and a combination of moderate review length and positive product evaluation statements usually led to good helpfulness ratings [Schindler and Bickart 2012].

In addition to inherent bias in review itself, the voting outcome may suffer from bias caused by self-selected behavior of consumers [Hu, Zhang et al. 2009, Chen, Zheng et al. 2013].

Since a review usually reflects its authors' gender, ethnicity, and income, as well as other social demographic attributes [Li and Hitt 2008], the same review could be very appealing to consumers who share similar attributes with author, but not appealing at all to others [Wang, Zhang et al. 2010]. This made the helpfulness votes received by a review more likely to be casted by those preferred consumers, led to aggregation bias.

Some evidence has been found in online review research. For example, in a study of reviews about more than 4 millions books, CDs, and videos on amazon.com, it was found that the accompanied product rating, which was submitted together with review by reviewer, was following the J-shaped distribution [Hu, Zhang et al. 2009]. A J-shaped distribution indicates those ratings (1 to 5) a product receives do not follow normal distribution. Instead, they tend to cluster around the extremely high range like 5 and 4 and the extremely low range, like 1 and 2. A recent study also found that online reviews with extreme opinions usually received more helpfulness votes than reviews with neutral or mixed opinions [Cao, Duan et al. 2011].

Aggregation bias is related to concomitant bias, also called under-reporting bias—that is consumers with different perspectives on a review's helpfulness tend to have different willingness to vote. In other words, there are many consumers who have read the review and also have an opinion on the review about its helpfulness but may choose not to cast their vote. This phenomenon was first observed and systematically summarized by Bradley Horowitz, former Yahoo VP of Advanced Development. He found that in terms of contribution to user-generated content, out of every 100 consumers, only 1 would contribute new content, like an online review; only 10 will vote on the helpfulness or usefulness of that content; and the other 89 consumers were merely content users or observers, hence he called this 1:10:89 rule. This phenomenon was also explained in more detail later in popular practitioner books [McConnell and Huba 2007, Howe 2008].

The combined effect of aggregation bias and concomitant bias may lead to a disproportional higher percentage of positive votes on review helpfulness because many consumers tend to be more conservative in terms of giving explicit negative ratings. For example, 99% of eBay user ratings were positive, not because they were all good sellers, but because buyers tend to withhold bad ratings or solve the problem with sellers in a different way [Dellarocas and Wood 2008]. In a recent study about movie ratings given by blockbusters consumers, it was found that the average rating a movie received was higher than they actually were after adjustment [Chen, Zheng et al. 2013]. Thus, the ratio of Yes votes received by an amazon.com review could be inflated due to self-selected behavior, so we have the following hypothesis:

Hypothesis 1: In a non-manipulated context, the helpfulness rating for the most helpful reviews of a product on amazon.com is higher than a rating from voters who were sampled from randomly selected population.

Next, we explore the resistance to manipulation by online review helpfulness votes.

1.3 The Resistance to Manipulation

Existing literature indicates when consumers vote on review helpfulness, their decisions were not based on review content only, but also influenced by contextual information, including existing helpfulness rating, which is explicitly displayed on top of the review [Baek, Ahn et al. 2012]. Actually, according to the Accessibility-Diagnosticity framework [Menon, Raghubir et al. 1995], existing helpfulness ratings are important diagnostic input with task-specific and goal-specific indicators for consumers to choose a review and give his or her own helpfulness vote on the review.

Tversky and Kahneman [1974] found that in many circumstances when people have to make trivial decisions, they tend to use heuristics and start from an initial value and then come up the final outcome based on it. Previous research also found that information that is activated to solve a comparative anchoring task would subsequently be more accessible when individuals make absolute judgments [Strack and Mussweiler 1997]. Thus, the more accessible a piece of anchoring information, such as existing helpfulness ratings, the more likely it becomes a starting point for an individual to estimate and vote on the helpfulness of related reviews.

Because of this anchoring effect, though online shoppers would estimate the helpfulness by making adjustments from anchoring value, most of the time, the adjustments are insufficient and lead to estimation closer to anchor value [Epley and Gilovich 2001]. Thus, existing helpfulness ratings could influence consumers' estimations on review helpfulness and bias the new vote towards it.

However, though the anchoring effect is quite robust, it could be mitigated by anchor-inconsistent knowledge [Mussweiler, Strack et al. 2000]. This led to two possible outcome scenarios. In the first scenario, when a consumer was initially primed by the existing helpfulness rating and then found the review was significantly less helpful than the current helpfulness rating indicated, the counter-effect generated by anchor-inconsistent knowledge becomes dominate; his expectation falls short, and he may vote down the helpfulness rating. In the second scenario, when an existing helpfulness rating is very low but the shopper felt its helpfulness far exceeds the exiting rating indicates, he may vote up the review—again opposing the anchoring effect.

Thus, the degree of inconsistency between the consumer's perception and the current rating plays a key role in the fluctuation of review rankings. If the inconsistency is large enough, it could overcome the anchoring effect currently working on the consumer and motivate the shopper to vote against the existing rating. If it is not significant, the anchoring effect would dominate and influence the vote towards the current rating. The underlying cognitive mechanism is not dissimilar to those described in other research domains, such as the disconfirmed expectations theory [Oliver 1977] in marketing, expectation/disillusion theory [Sigelman and Knight 1983] in political science, and expectation confirmation theory [Bhattacherjee 2001] in information systems.

Most inconsistency comes from manipulations. Recently, manipulated product reviews become serious challenge to online retailers, because the authenticity of online reviews is questioned [Moyer 2010]. It is very likely that consumers suspect some glowing reviews are posted by hired guns, though they could not confirm that. In fact, though 2011 Social Shopping Study[1] ranks amazon.com as the top most credible peer-review source (63%)—followed by independent review sites like epinion.com (51%), search engines (50%), and social networking sites (SNS)—it is still not trusted by at least one third of consumers (37%). When a helpfulness rating was deliberately manipulated, we may observe temporary significant inconsistence between the actual quality of the review and its helpfulness ratings. For example, some manufacturers who want to boost the reputation of their product may manipulate online reviews and the helpfulness ratings of such reviews [Dellarocas 2010], which could make a review less helpful than it appears, thus increasing the inconsistency.

As indicated previously, the most helpfulness reviews are the prime target for manipulation via helpfulness vote. However, the resulting effect between *most favorable* and *most critical* reviews could be different.

For *most favorable reviews,* though consumers may be aware that some of them were manipulated, they may not be able to identify them. This is because though a favorable review could deliberately emphasize the positive features of a product and only cover minor negative issues, or not covering the latter at all, such reviews are usually subjective and difficult for consumers to evaluate its authenticity (e.g. a manipulated review could explain in details the good feelings of using a product, but there is no way to validate whether such feeling is authentic). Thus, a consumer is likely to be more influenced by its current helpfulness rating even there is inconsistency. So the anchoring effect of existing helpfulness ratings may be dominant in voting decisions for the most helpful favorable reviews.

In contrast, critical or negative reviews are deemed relatively more objectively, authentic, and important because of negativity bias [Rozin and Royzman 2001].

Kahneman and Tversky [1979] found that people have a natural tendency of risk aversion; they put more weight on loss than on gain in making decisions. Such tendency leads people to pay more attention and give more weight to negative rather than positive information, because the former may reveal potential risk. Negativity bias is innate with our nature and was observed even on 3-month-old infants [Kiley Hamlin, Wynn et al. 2010]. Since most critical reviews usually reveal potential risk aspects of a product, they are regarded more important than favorable reviews. Critical reviews are also less available than favorable or positive reviews for a product. According to Pinch and Kesler [2011] only 1 out of 10 reviews from consumers were critical or negative.

Because of negativity bias, risk aversion and scarcity of critical reviews, consumers tend to give more weight on critical review content than their existing helpfulness ratings. If there is a large inconsistency between existing ratings and a consumer's assessment, the counter anchoring effect could dominate the anchoring effect. As a result, he or she may disagree with the existing helpfulness rating for the review and vote against it accordingly. Thus we have:

Hypothesis 2: The inconsistency between consumers' assessment of a review's helpfulness and its existing helpfulness rating has a stronger influence on voting outcomes for most helpful critical reviews than for most helpful favorable reviews.

[1] For details, please refer to: http://www.powerreviews.com/assets/download/Social_Shopping_2011_Brief1.pdf

2. Research Method

We designed two experiments to verify the above two hypotheses. Experiment 1 focused on detecting the general reliability hypothesis. We randomly selected a group of online shoppers—in this case a convenient sample of graduate students—and let each subject rated the helpfulness of the most helpful reviews without revealing the actual helpfulness rating each review already received. Then we compared the aggregated results with the rating the review received from amazon.com. If there were significant differences, the hypothesis would be confirmed.

Experiment 2 used the same convenient sample and focused on detecting the relative strength of the anchoring effect of existing helpfulness ratings and the counter effect of inconsistency with consumers' own assessments. We either raised or lowered the actual helpfulness rating a most helpful review receives. We then asked each subject to rate the review when they were primed by manipulated ratings and to decide whether the review was helpful or not. We also included the same reviews without any helpfulness-rating information as a benchmark.

3. Experiment 1

In Experiment 1, we first collected a sample of most helpful reviews from top-selling products shown in the top page of selected product categories at amazon.com, screen-captured the review, and masked their helpfulness rating information. Then, we asked subjects to read those reviews and vote on the helpfulness for each of them. The results were then compared with the original ratings.

To select the most helpful reviews, we identified the top three best-selling products from search, experience, and credence (SEC) product categories, respectively, to cancel out potential bias originated from product categories [Nelson 1970, Darby and Kami 1973, Nelson 1974]. Search products refer to those commodities for which a consumer could evaluate their quality before making the purchase. All those mass-produced products and groceries belong to this category [Nelson 1970]. Experience products are those for which a consumer can only evaluate its quality after the purchase or use of it. For example, a consumer cannot evaluate the quality of food and service offered by a new local restaurant until he visits it and has a dinner there [Nelson 1974]. Credence products are those for which, even after a prolonged period of use, a consumer may still not be able to evaluate their quality. Many service products like car maintenance and insurance belong in this category [Darby and Kami 1973]. We also need to note that some products or services had attributes in all SEC categories, and we classified them into a specific category according to the attributes with which consumers were most concerned. Thus, one product may fall into different SEC categories for different consumers, depending on the latter's knowledge and experience. For example, an experienced car mechanic may have little difficulty figuring out the maintenance quality offered by his peers by just driving the car out of the auto-shop.

Online review could help consumers evaluate product quality before they make purchase decisions. However, its impact on the purchase decision-making process is different across SEC categories. For a search or experience product, since consumers could easily evaluate the validity of the review after receiving the product, many would assume those reviews should have been authenticated. However, that may not be the case for credence products.

In this experience, we choose reviews from all three categories to help us compare findings more comprehensively and avoid differences caused by SEC.

Table 1. Survey products selection

Search	• T-Mobile G2x 4G Android Phone • Canon PowerShot A30000IS • HP LaserJet Pro P1102w Printer
Experience	• "LUNGS" (Audio CD) • "Heritage" (Audio CD) • "L.A. Noire" (Video Game)
Credence	• OxyElite Pro • Acidophilus Pearls • Whey

For each product, we chose the top three most helpful reviews given by amazon.com. We collected a total of 27 most helpful reviews. We then ranked them based on the total number of votes each review received. We selected 9 reviews from them to ensure there were at least three reviews from each SEC category. In addition, the total number of votes each review received was spread out. They ranged from as low as 0 to 10 to as high as 90 or higher. The 9 identified product reviews were screen-captured from the website to retain the original format and style. By doing this, we retained the same review style and format as on amazon.com to eliminate other potential distraction factors.

The helpfulness rating for each review was recorded and then masked. Each review was inserted and coded into one web page with a simple "helpful or not" question under it. About 80 graduate students from four classes in a Midwestern university were invited to participate in the experiment as a convenient sample of randomly selected shoppers. There was no monetary compensation, though an optional extra credit towards participants' final grades was given for participation. Each participant was asked to read all 9 reviews, and for each review, they were asked to vote *Yes* or *No* on its helpfulness. Demographic information of the participants, including gender, age, income, etc., was collected at the beginning of the experiment. A total of 74 valid responses were received and the descriptive statistics of the subjects is in table 2. Those 6 incomplete responses were discarded.

Table 2 : Descriptive statistics of subjects

Attributes	Descriptive Statistics
Gender	Female: 31.1% Male: 68.9%
Age	20~29: 48.6% 30~39: 45.9% 40~49: 2.7% 50~59: 2.7%
Ethnicity	African: 17.6% Asian/Pacific Islander: 39.2% Caucasian: 36.5% Hispanic: 6.8%
Income	<$20K: 16.2% $20 ~39K: 23% $40~59K:25.7% >$60K: 36.5%

3.1 Data Analysis

We compared the helpfulness ratings at amazon.com with those obtained from the 74 valid responses. The helpfulness rating was calculated using the following formula:

$$helpfulness\ rating = \frac{total\ number\ of\ YES\ votes\ received}{total\ votes\ received}$$

The total votes received included both *Yes* and *No* voting. We used the binomial test to compare the amazon.com ratings with the sample helpfulness ratings. We regarded the existing helpfulness ratings on amazon.com as benchmarks and then tested whether the rating we collected from subjects was statistically similar to the benchmark. In Table 3, the "Yes" column and "Total" column under both Ratings categories refer to the votes a product review received. The "Rating" column is the helpfulness rating indicated above. Table 3 lists the *p*-values in the right-most column regarding the difference between the helpfulness ratings at amazon.com and those obtained from the survey participants. All these p-values are less than 0.05, and all sampled ratings are lower than ratings on amazon.com. *Thus, H1 is supported.*

Table 3: Summary differences between amazon.com and sample survey ratings

SEC	**Product Name**	**Amazon.com Ratings**			**Sample Ratings**			**Symp. Sig (1-tail)**
		Yes	**Total**	**Rating**	**Yes**	**Total**	**Rating**	
Search	T-Mobile G2x 4G Android Phone	13	15	*0.87*	52	74	*0.70*	.000
Search	Canon PowerShot A3000IS	53	57	*0.93*	54	74	*0.73*	.000
Search	HP LaserJet Pro P1102w Printer	61	64	*0.95*	52	74	*0.70*	.000
Experience	LUNGS (Audio CD)	36	39	*0.92*	37	74	*0.50*	.000
Experience	Heritage (Audio CD)	8	8	*1.00*	53	74	*0.72*	.000
Experience	L.A. Noire (Video Game)	23	26	*0.88*	14	74	*0.19*	.000
Credence	OxyElite Pro	82	92	*0.89*	30	74	*0.41*	.000
Credence	Acidophilus Pearls	47	47	*1.00*	54	74	*0.73*	.000
Credence	Whey	64	73	*0.88*	57	74	*0.77*	.006

A common rule of thumb to approximate a binomial distribution with the normal distribution is under the condition that both *np* and *n(1–p)* (*n*: sample size) are equal or greater than 10 [Gravetter and Wallnau 2009]. In our case, the smaller of these values did not exceed 9.62. Therefore, we cannot approximate normal distribution and there is no z-score report.

In addition to H1, we also found that all other conditions being equal, ratings with larger total votes bases, or voted on by a larger number of online shoppers, were closer to ratings given by participants in our experiment, which indicated the more total votes a most helpful review receives, the closer its helpfulness rating was to its true helpfulness value.

We used *chi*-square tests to detect any statistical differences on the helpfulness ratings based on four profile dimensions of survey participants. These dimensions are: *gender, age, ethnicity,* and *income*. Given the fact that some sample cells are expected to contain less than 5 subjects, we applied the Monte Carlo simulation with the sample size of 10,000 for those cases. The results are summarized in Table 4. It contains *p*-values of the *chi*-square tests regarding the user review helpfulness ratings based on gender, age, ethnicity and income for the online shopping of the survey participants (the blank cells mean p ≥ .10, or not significant). Provided the small sample size, we use α = .10 as the cutoff point.

According to the results, there are at least some significant differences based on one or a combination of two or more profile attributes of online shoppers for the search and experience goods, except for the Android phone. Gender was a significant factor for two search goods: the Canon PowerShot digital camera and HP LaserJet printer. The digital camera had some differences in the ratings among different genders and ethnicities. The laser printer saw the different ratings by gender and income levels. For experience goods, gender and age were not significant factors. Ethnicity was a factor for one audio CD. Finally, income was a factor for the video game. In contrast, we do not see any significant differences for the credence goods.

Table 4: Demographic influences on helpfulness ratings

SEC	Product Name	Gender	Age	Ethnicity	Income	Mobile Use
Search	T-Mobile G2x 4G Android Phone					
Search	Canon PowerShot A3000IS	.069		.094*		
Search	HP LaserJet Pro P1102w Printer	.082			.017*	
Experience	LUNGS (Audio CD)					.028
Experience	Heritage (Audio CD)			.063*		
Experience	L.A. Noire (Video Game)				.060	
Credence	OxyElite Pro					
Credence	Acidophilus Pearls					
Credence	Whey					
	Pearson Chi-Square, df	1	3	3	3	4

*: Based on Monte Carlo simulation (10,000 samples)

In the previous experiment, we identified the significant difference between helpfulness ratings displayed on amazon.com and those given by randomly selected populations. We also found that the former is significantly higher than the latter. Thus it is very likely that the most favorable review on amazon.com we collected has inflated helpfulness ratings, though it is likely that they would gradually deflate when the total number of votes they receive increases. In addition, we found the demographic attributes of consumers could significantly influence their voting outcome, though the specific impact is not consistent across product categories. Further study is needed to find out hidden patterns of such influence.

Next, we explore H2, *the resistance to manipulation by helpfulness votes.*

4. Experiment 2

In Experiment 2, we chose reviews from the vitamin and supplements product category as the sample reviews for testing. These products are most commonly purchased across consumers with different incomes, educational backgrounds, ethnicity and gender. They are also regarded as credence goods whose product quality cannot be determined even after purchase [Darby and Karni 1973]. In other words, we as consumers have to largely depend on other consumers' reviews to evaluate the product and make purchase decisions. Thus, how a review that describes one's experience and knowledge on credence product may play a more critical role for making purchase decision by other consumers [Bae and Lee 2011].

We selected the most helpful reviews for the top 2 best-selling products in the vitamin and supplements category. One most helpful favorable review (Review 1) and one most helpful critical review (Review 2) were identified.

To setup the manipulation scenarios, each review was adapted with one of three helpfulness rating treatment conditions: *a higher rating similar to or above what amazon.com was showing*, *no rating to be displayed*, and *a significantly lower rating than amazon.com was showing*. For the favorable and critical reviews, the three manipulation ratings are (100%, none, 46%) and (79%, none, 45%) respectively. A treatment for review 1 was combined with a treatment for review 2 as one scenario. There were three scenarios being created. We recruited 108 subjects from two universities in the Midwestern and Southwestern United States (male 66.7%, female 33.3%; age 20-29 54.6%, 30-39 26.9%, 40-49 13% and 50-59 0.9%). One extra credit was given as incentive. Each participant was asked to read reviews randomly chosen from one scenario and vote for its helpfulness. The three scenario profiles, including higher and lower rating configurations, and the number of participants participated for each scenario (N) were:

- *Scenario 1: higher rating* for Review 1(100%) and *no rating* for Review 2 (N=37)
- *Scenario 2: no rating* for Review 1 and *higher rating* for Review 2 (79%) (N=33)
- *Scenario 3: lower rating* for Reviews 1(46%) and 2(45%) (N=38)

4.1 Data Analysis

We used binomial tests on the dataset, because the dependent variable is binary (*Yes* or *No*) and the sample sizes were relatively small. However, the statistical significances are identical if we use Z approximations; the sample sizes were not to meet the binary-test conditions we used for Experiment 1. The results of Experiment 2 were summarized in Table 4. Consumers were generally influenced by the manipulated helpfulness ratings with interesting differences between the most favorable and critical reviews.

Table 5: Summary of helpfulness ratings given by subjects for review 1 and 2

	Favorable Review	**Critical Review**
Higher rating	60.0% (100% shown)*	73.7% (79% shown)
No rating	57.9% (no rating shown)	78.8% (no rating shown)
Lower rating	54.4% (46% shown)	87.5% (45% shown)*

*. Significant at the 0.05 level (1-tailed)

For favorable reviews, the influence of the adjustment and anchoring effects was consistent in all three scenarios. Though only 60% of participants exposed in a 100% existing helpfulness rating context thought the review was helpful, it was still significantly higher than the “no rating” and “lower rating” contexts. Meanwhile, the helpfulness rating given by participants in the lower rating context (46%) was less than both “higher” and “no rating” contexts. We also found evidence of the counter-effect due to different degrees of inconsistency, though it was not as strong as the anchoring effect in the most favorable review. For example, if we regard the “no rating” scenario as the benchmark (or actual rating) the review deserves (57.9% in this case), the inflated helpfulness rating (100%), which had a larger inconsistency compared to the benchmark, led to a 40% correction in participant votes (decreased to 60%). This was much larger than the 8.4% correction the deflated helpfulness rating received (increased to 54.4%), probably because the deflated rating has a much smaller inconsistency (46%) to benchmark.

For critical review, we observed a comparatively larger influence of counter-effect than anchoring effect. When the deflated helpfulness rating (45%) was presented to participants, the voting outcome by the participant was not just increased to a level that was more than the benchmark rating (78.8%) but also significantly exceeded it (87.5%). This indicated that the anchoring effect of a deflated rating was less dominant than its counter-effect due to inconsistency and general risk aversion by consumers, as well as consumers’ increased trust of the authenticity of critical reviews mentioned before. *Thus, H2 is also supported*.

5. Implications, Limitations and Future Research

Existing literature found in online reviews can be manipulated and biased in many ways [Mackiewicz 2007]. The helpfulness voting practice adopted by amazon.com and other online retailers is used to mitigate such challenges like sequential bias. However, our study indicated the helpfulness rating information may not be reliable and is generally inflated for those most helpful reviews. Considering the helpfulness ratings were being used by many researchers, such as Mudambi and Schuff [2010], as review helpfulness benchmarks, we should be cautious when using the ratings as benchmark evidences. Since such inflation tends to decrease as more votes come in, certain adjustment techniques could be used to make the rating closer to its true value and this study could serve as a starting point.

This study also sheds insight on the impacts of helpfulness rating manipulation. According to website traffic analysis service compete.com, amazon.com attracts 70 to 80 million visitors each month. Rating manipulations, even for an hour, can influence a large number of shoppers, possibly resulting in substantial financial gains for manufacturers or vendors. Our findings indicate that trying to manipulate helpfulness ratings may not be cost effective; it could even backfire. An inflated helpfulness rating for a most favorable review does not necessarily lead to consumer trust and product sales. On the other side, when a vendor’s product is being critically or negatively reviewed on amazon.com and such critical reviews are being considered very helpful, if the vendor tries to manipulate the rating and bust it down, consumers might feel those critical reviews are more helpful than they really are, thereby increasing the helpfulness ratings.

One major limitation of this research is the small number of reviews being used in this study. With a limited number of reviews as a sample, we could not exclude many context variables that might affect the helpfulness vote outcome, which may have led to the differences we identified in this study. Another limitation is the convenient sample of graduate students we used. There might be inherent bias in the student population, even though the graduate student group probably had less of such bias compared with undergraduates.

Future research could focus on further expanding the review samples and testing the extent of our findings in different product categories and domains. Another direction is how to improve the rating process so as to make it more objective and less influenced by self-selected behavior. It would also be interesting to use some statistical methods to adjust the rating [Chen, Zheng et al. 2013].

This exploratory study does not intend to assert that reviews with manipulated helpfulness votes always affect consumers exactly the same way we saw in this paper. However, it is a common misconception that we as consumers vote without being influenced by existing ratings, at least for the two best-selling vitamin products (credence goods) we used in this study. Future studies can investigate the validity of this finding using other types of popular products.

6. Conclusion

In this paper, we explored the reliability of online review helpfulness ratings on amazon.com. We found the ratings for those most helpful reviews are consistently inflated compared with the ratings provided by our sampling subjects. We also explored the anchoring effect of existing ratings as well as its counter-effect due to inconsistency between online shopper expectations. We found that the inconsistency between online shoppers' own assessment of a review's helpfulness and its existing helpfulness rating has a stronger influence on voting outcomes for top critical reviews than for top favorable reviews. Our findings have important implications for both academic research in online review and online retailing practitioners.

REFERENCES

Bae, S. and T. Lee, "Product type and consumers' perception of online consumer reviews." *Electronic Markets* Vol. 21, No. 4: 255-266, 2011.

Baek, H., J. Ahn and Y. Choi, "Helpfulness of Online Consumer Reviews: Readers' Objectives and Review Cues." *International Journal of Electronic Commerce* Vol. 17, No. 2: 99-126, 2012.

Banks, S. , "*Putting Yelp in their rear-review mirror.*" Los Angeles Times, 2013

Bhattacherjee, A., "Understanding Information Systems Continuance: An Expectation-Confirmation Model." *MIS Quarterly* Vol. 25, No. 3: 351-370, 2001.

Brian Arthur, W., Y. M. Ermoliev and Y. M. Kaniovski, "Path-dependent processes and the emergence of macro-structure." *European Journal of Operational Research* Vol. 30, No. 3: 294-303, 1987.

Cao, Q., W. Duan and Q. Gan, "Exploring determinants of voting for the "helpfulness" of online user reviews: A text mining approach." *Decision Support Systems* Vol. 50, No. 2: 511-521, 2011.

Chappell, B., "Online Review-Rigging Firms To Pay Fines In Yogurt Shop Sting." Retrieved September 23, 2013, from http://www.npr.org/blogs/thetwo-way/2013/09/23/225459295/online-review-rigging-firms-to-pay-fines-in-yogurt-shop-sting.

Chen, H., E. Zheng and Y. Ceran, "The Power of Silence: An Analysis of the Aggregation and Reporting Biases in User-Generated Contents." *Working Paper*

Chen, Y. and J. Xie, "Online Consumer Review: Word-of-Mouth as a New Element of Marketing Communication Mix." *Management Science* Vol. 54, No. 3: 477-491, 2008.

Clemons, E., G. Gao and L. Hitt, "When Online Reviews Meet Hyperdifferentiation: A Study of the Craft Beer Industry." *J. Manage. Inf. Syst.* Vol. 23, No. 2: 149-171, 2006.

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

Cui, G., H.-K. Lui and X. Guo, "The Effect of Online Consumer Reviews on New Product Sales." *International Journal of Electronic Commerce* Vol. 17, No. 1: 39-58, 2012.

Darby, M. R. and E. Kami, "Free Competition and the Optimal Amount of Fraud." *Journal of Law and Economics* Vol. 16, No.: 66-86, 1973.

Dellarocas, C., "Online reputation systems: how to design one that does what you need." *MIT Sloan Management Review* Vol. 51, No. 3: 33-38, 2010.

Dellarocas, C. and C. A. Wood, "The sound of silence in online feedback: estimating trading risks in the presence of reporting bias." *Management Science* Vol. 54, No. 3: 460-476, 2008.

Epley, N. and T. Gilovich, "Putting Adjustment Back in the Anchoring and Adjustment Heuristic: Differential Processing of Self-Generated and Experimenter-Provided Anchors." *Psychological Science* Vol. 12, No. 5: 391-396, 2001.

Ghose, A. and P. G. Ipeirotis, "Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics." *Knowledge and Data Engineering, IEEE Transactions on* Vol. 23, No. 10: 1498-1512, 2011.

Gravetter, F. J. and L. B. Wallnau, "*Statistics for the Behavioral Sciences.*" Belmont, CA, Wadsworth, 2009

Howe, J., "*Crowdsourcing: How the power of the crowd is driving the future of business*," Random House, 2008

Hu, N., J. Zhang and P. A. Pavlou, "Overcoming the J-shaped distribution of product reviews." *Communications of the ACM* Vol. 52, No. 10: 144-147, 2009.

Kahneman, D. and A. Tversky, "Prospect Theory: An Analysis of Decision under Risk." *Econometrica* Vol. 47, No. 2: 263-291, 1979.

Kapoor, G. and S. Piramuthu, "Sequential Bias in Online Product Reviews." *Journal of Organizational Computing & Electronic Commerce* Vol. 19, No. 2: 85-95, 2009.

Kiley Hamlin, J., K. Wynn and P. Bloom, "Three-month-olds show a negativity bias in their social evaluations." *Developmental science* Vol. 13, No. 6: 923-929, 2010.

Korfiatis, N., E. García-Bariocanal and S. Sánchez-Alonso, "Evaluating content quality and helpfulness of online product reviews: The interplay of review helpfulness vs. review content." *Electronic Commerce Research and Applications* Vol. 11, No. 3: 205-217, 2012.

Li, X. and L. M. Hitt, "Self-Selection and Information Role of Online Product Reviews." *Information systems research* Vol. 19, No. 4: 456-474, 2008.

Mackiewicz, J., "*Reviewer bias and credibility in online reviews*." 72nd Annual Convention of the Association for Business Communication. Washington, D. C., 2007

Maes, P., "Agents that reduce work and information overload." *Communications of the ACM* Vol. 37, No. 7: 30-40, 1994.

McConnell, B. and J. Huba, "*Citizen marketers: When people are the message*," Kaplan Publication, 2007

Menon, G., P. Raghubir and N. Schwarz, "Behavioral Frequency Judgments: An Accessibility-Diagnosticity Framework." *Journal of Consumer Research* Vol. 22, No. 2: 212-228, 1995.

Moyer, M., "Manipulation of the Crowd." *Scientific American Magazine* Vol. 303, No. 1: 26-28, 2010.

Mudambi, S. and D. Schuff, "What Makes A Helpful Online Review? A Study of Customer Reviews on Amazon.com." *MIS Quarterly* Vol. 34, No. 1: 185-200, 2010.

Mukherjee, A., B. Liu and N. Glance, "*Spotting fake reviewer groups in consumer reviews*." Proceedings of the 21st international conference on World Wide Web, ACM, 2012

Mussweiler, T., F. Strack and T. Pfeiffer, "Overcoming the inevitable anchoring effect: Considering the opposite compensates for selective accessibility." *Personality and Social Psychology Bulletin* Vol. 26, No. 9: 1142-1150, 2000.

Nelson, P., "Information and consumer behavior." *Journal of Political Economy* Vol. 78, No. 2: 311-329, 1970.

Nelson, P., "Advertising as information." *Journal of Political Economy* Vol. 82, No. 4: 729-754, 1974.

Oliver, R. L., "Effect of Expectation and Discontinuation on Postexposure Product Evaluations: An Alternative Interpretation." *Journal of Applied Psychology* Vol. 62, No. 4: 480-486, 1977.

Pan, Y. and J. Q. Zhang, "Born unequal: a study of the helpfulness of user-generated product reviews." *Journal of Retailing* Vol. 87, No. 4: 598-612, 2011.

Park, D.-H., J. Lee and I. Han, "The Effect of On-Line Consumer Reviews on Consumer Purchasing Intention: The Moderating Role of Involvement." *International Journal of Electronic Commerce* Vol. 11, No. 4: 125-148, 2007.

Pinch, T. and F. Kesler, "How Aunt Ammy Gets Her Free Lunch: A Study of the Top-Thousand Customer",2011. Reviewers at Amazon.com, Available at http://www.freelunch.me/filecabinet.

Purnawirawan, N., N. Dens and P. De Pelsmacker, "Balance and Sequence in Online Reviews: The Wrap Effect." *International Journal of Electronic Commerce* Vol. 17, No. 2: 71-98, 2012.

Rozin, P. and E. B. Royzman, "Negativity bias, negativity dominance, and contagion." *Personality and social psychology review* Vol. 5, No. 4: 296-320, 2001.

Salganik, M. J., P. S. Dodds and D. J. Watts, "Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market." *Science* Vol. 311, No. 5762: 3, 2006.

Schindler, R. M. and B. Bickart, "Perceived helpfulness of online consumer reviews: the role of message content and style." *Journal of Consumer Behaviour* Vol. 11, No. 3: 234-243, 2012.

Sigelman, L. and K. Knight, "Why Does Presidential Popularity Decline? A Test of the Expectation/Disillusion Theory." *The Public Opinion Quarterly* Vol. 47, No. 3: 310-324, 1983.

Strack, F. and T. Mussweiler, "Explaining the enigmatic anchoring effect: Mechanisms of selective accessibility." *Journal of Personality and Social Psychology* Vol. 73, No.: 437-446, 1997.

Tversky, A. and D. Kahneman, "Judgment under Uncertainty: Heuristics and Biases." *Science* Vol. 185, No. 4157: 1124-1131, 1974.

Wan, Y., S. Menon and A. Ramaprasad, "A Classification of Product Comparison Agents." *Communications of the ACM* Vol. 50, No. 8: 65-71, 2007.

Wang, C., X. Zhang and I.-H. Hann, "*Social bias in online product ratings: a quasi-experimental analysis.*" WISE 2010. St. Louis, MO., 2010

Ye, Q., R. Law, B. Gu and W. Chen, "The influence of user-generated content on traveler behavior: An empirical investigation on the effects of e-word-of-mouth to hotel online bookings." *Computers in Human Behavior* Vol. 27, No. 2: 634-639, 2011.

Zhu, F. and X. Zhang, "Impact of Online Consumer Reviews on Sales: The Moderating Role of Product and Consumer Characteristics." *Journal of Marketing* Vol. 74, No. 2: 133-148, 2010.