5-2012

An Exploratory Study: “Blind-Testing” Consumers How They Rate Helpfulness of Online Reviews

Makoto Nakayama
DePaul University, mnakayama@cdm.depaul.edu

Yun Wan
University of Houston - Victoria, wany@uhv.edu

An Exploratory Study: “Blind-Testing” Consumers How They Rate Helpfulness of Online Reviews

Makoto Nakayama
DePaul University
mnakayama@cdm.depaul.edu

Yun Wan
University of Houston - Victoria
wany@uhv.edu

Abstract

Amazon.com uses the helpfulness vote (H-Vote) on consumer product reviews to signal quality. The top reviews with the best H-Vote are thus prominently displayed so that consumers readily use them for their purchase decision-making. However, the influence of those H-Votes may not be the same depending on contexts. We conducted a pilot survey experiment by using 108 consumers to investigate. The results showed that consumers were not swayed by a positive or negative H-Vote for a top *favorable* review in Amazon.com. However, consumers significantly voted positively for a top *unfavorable* (or critical) review in Amazon.com when a *negatively-biased H-Vote* was shown. To further explore the underlying factors, we deconstructed the YES/NO-based H-Vote with a Likert-scale-based helpfulness rating for (a) product learning and (b) purchase decision-making. While the difference in H-Vote did not affect the *learning* rating, it did impact the *decision* rating under a false-negative H-Vote. Implications and future research directions are discussed.

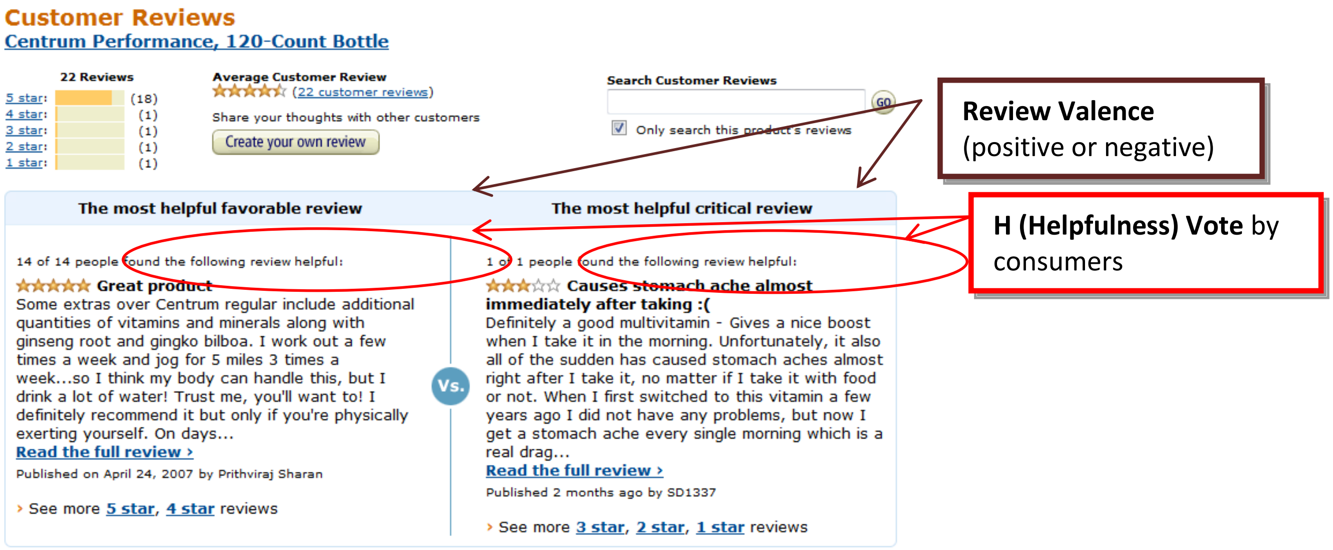
Keywords

Purchase Decision-Making, Helpfulness Vote (H-Vote), Framing Effect, Guttman (Binary) Scale, Likert Scale, False Positive.

1. Introduction

Peer-generated online reviews (*online reviews* or just *reviews* hereafter) can offer greater value to customers (Mudambi & Schuff, 2010). Empirical studies done in the past ten years report these reviews have strong influences on product sales (Zhu & Zhang, 2010).

According to the 2011 Social Shopping Study (PowerReviews, 2011), “one in two respondents spent 75% of their overall shopping time researching product as compared to just 21% in 2010.” The same study also indicates that online reviews at retailer websites have a greater impact than online community forums and social network sites (SNS), such as Facebook and Twitter. Among major retailer websites, Amazon.com is the most popular site for reviews (*ibid*.).



Customer Reviews

Centrum Performance, 120-Count Bottle

22 Reviews

5 star:	(18)
4 star:	(1)
3 star:	(1)
2 star:	(1)
1 star:	(1)

Average Customer Review
★★★★☆ (22 customer reviews)

Share your thoughts with other customers

Search Customer Reviews

☑ Only search this product's reviews

The most helpful favorable review

14 of 14 people found the following review helpful:

★★★★★ **Great product**
Some extras over Centrum regular include additional quantities of vitamins and minerals along with ginseng root and gingko bilboa. I work out a few times a week and jog for 5 miles 3 times a week...so I think my body can handle this, but I drink a lot of water! Trust me, you'll want to! I definitely recommend it but only if you're physically exerting yourself. On days...
Read the full review ›
Published on April 24, 2007 by Prithviraj Sharan

› See more 5 star, 4 star reviews

The most helpful critical review

1 of 1 people found the following review helpful:

★★★☆☆ **Causes stomach ache almost immediately after taking :(**
Definitely a good multivitamin - Gives a nice boost when I take it in the morning. Unfortunately, it also all of the sudden has caused stomach aches almost right after I take it, no matter if I take it with food or not. When I first switched to this vitamin a few years ago I did not have any problems, but now I get a stomach ache every single morning which is a real drag...
Read the full review ›
Published 2 months ago by SD1337

› See more 3 star, 2 star, 1 star reviews

Figure 1: Prominently displayed most favorable and critical reviews at Amazon.com

However, online reviews are known to be frequently manipulated (Hu, Liu, & Sambamurthy, 2011) and questioned for their authenticity (Moyer, 2010). For example, a manufacturer may post favorable reviews with more "YES" votes on its product in an attempt to increase product sales. It may also do the opposite on its opponents' products. Amazon.com attracts 70 to 80 million visitors each month[1]. Rating manipulations even for a day can influence a large number of shoppers, possibly resulting in substantial financial gains or losses for manufacturers.

This study focuses on the dynamics of helpfulness votes and their impacts on those prominently shown favorable and critical reviews at Amazon.com (Figure 1). We call them *top reviews* in this paper. The research question is *how existing helpfulness votes influence consumers when they cast their own vote, which is supposed to be their independent, unbiased judgment.* We also explore the impact of existing biased votes on two key sub-dimensions of helpfulness votes – the review value for (1) product learning as well as (2) purchase decision-making.

The following sections present theoretical background and hypotheses, method, results, implications and future research agenda, and conclusion.

2. Theoretical Background and Hypotheses

Review ranking based on votes is influenced by voting psychology. If consumers feel the review is less helpful than the current helpfulness votes indicate, their expectation falls short and they may vote down the helpfulness rating. On the other hand, they might feel the helpfulness of the review exceeds what the current votes indicate. Then, they vote the review up. Thus, the degree of discrepancy between consumers' perception and the current votes plays a key role in the fluctuation of review rankings. The underlying theory is not dissimilar to existing theories in other research domains, such as the disconfirmed expectations theory (Oliver, 1977) in marketing, expectation/disillusion theory (Sigelman & Knight, 1983) in political science, and expectation confirmation theory (Bhattacherjee, 2001) in information systems (IS). These

[1] http://siteanalytics.compete.com/amazon.com/?metric=uv

existing theories shed light on the basic dynamics of helpfulness voting. However, we still are not certain to what extent current biased votes impact new votes. In addition, such impacts may vary depending on the valence (favorable or critical) of the reviews.
Amazon reviews are not necessarily written by real buyers of the products. There are open or private solicitations of paid review contributions by product manufacturers or sellers to promote their product.[2] Consumers generally expect shills (writers of fake reviews) to write some reviews, in particular, overly favorable ones. While evidence is anecdotal, there were 29 news articles published between July 29 and August 28, 2011, when we searched Google News with the key words "fake reviews." Because of these frequent news reports on "shills," consumers are cautious about rave or booing comments that can be seen among top-rated reviews. Although the 2011 Social Shopping Study (PowerReviews, 2011) 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 over one third of consumers (37%). Thus, we expect consumers are more conservative about high H-Votes (since shills potentially inflate them) for those top favorable reviews and tend to use their own judgment to evaluate whether the review is helpful or not. This can be viewed as one form of the framing effect (Kahneman, 2003; Tversky & Kahneman, 1981) in which consumers are conditioned to discount the available information due to the prevailing perception of shills.

H1: For top favorable (positive) product reviews, the existing H-Vote does not affect how consumers cast their own vote on the review.

Most online product reviews are positive. The 2011 Social Shopping Study (ibid.) also reports the top 3 occasions of consumer online participations are (1) when consumers purchased a product, (2) when they liked a retailer, and (3) when they liked a manufacturer. According to Pinch and Kesler (2011), "88% of our respondents reported that they gave either 'all' or 'most' [favorable] reviews." This indicates only 1 out of 10 reviews from a consumer will be negative. In other words, positive reviews are much more abundant than negative reviews. Given that consumers are much less exposed to negative reviews than positive reviews, negative reviews and their H-Vote may have more impact on consumer perceptions than that of positive reviews. This is the reverse form of the framing effect we note for H1. Thus:

H2: For top critical (negative) product reviews, the existing H-Vote does affect how consumers cast their own vote on the review.

To further explore the psychology of H-Votes, we need to explore the underlying factors. Weiss, Lurie and MacInnis (2008) note that the underlying factors include the goal orientations of consumers. Consumers view the helpfulness of a review differently when they are researching a product (Learning) and when they are buying a product (Decision). Furthermore, the binary (YES/NO) nature of H-Votes may impact the rating outcomes. The Guttman (binary) scale may oversimplify the underlying phenomenon and may make statistical analyses more difficult (Smith, 1960). That is, H-Votes may be over-accentuating consumer preferences without considering consumer goal orientations. Thus we need to explore if H1 and H2 still hold water

[2] Examples are seen both online (http://www.getelastic.com/asking-for-customer-reviews/) and offline (http://www.searchinfluence.com/2010/04/buying-yelp-reviews/).

when we deconstruct the H-Vote into the two, Likert-scale sub-dimensions: (1) helpfulness for product learning (Learning) and (2) helpfulness for purchase decision-making (Decision).

On one hand, the learning aspect involves the knowledge accumulation over time beyond one review. On the other hand, the decision aspect motivates consumers to "narrow, support, and justify a specific course of action" (Weiss, Lurie, & MacInnis, 2008). Because making purchase decisions carries higher stakes and incurs financial commitment for the consumer side, we expect consumers will be more sensitive to H-Votes when they use them for purchase decision-making rather than merely researching the product.

H3: Consumers are more sensitive to the difference in H-Vote for purchase decision-making than for researching the product

3. Method

We used a stylized experiment randomly selecting 2 prominently displayed reviews from the first best-selling product page for vitamins and supplements at Amazon.com[3]. We showed these reviews to participants of this study and asked them to rate the helpfulness of the reviews. Each review was shown with one of three different treatment conditions on H-Vote: a higher H-Vote (close to the vote Amazon.com was showing), no H-Vote, and a lower H-Vote (significantly lower than what Amazon.com showing).

There are many online reviews available on the Web. In this pilot survey, we focused on reviews for vitamin supplements. These products are most commonly purchased by consumers with different incomes, educational backgrounds, ethnicity and gender. They are also regarded as credence goods whose product quality cannot be determined exactly even after purchase (Darby & Karni, 1973). In other words, we as consumers have to rely on other consumers' experience and perceptions. What and how a review describes one consumer's experience and knowledge may play a more critical role for purchase of credence goods than other types of goods. We randomly selected the most helpful favorable review for one product and the most helpful critical review for another product shown in the top category page at Amazon.com.

Survey volunteers were recruited with an extra credit for a course taught at two universities in the Midwest and Southwest. there were 108 participants in total. Each was randomly assigned to one of the three versions of the survey questionnaires. Following are the specific treatment and corresponding number of participants:

(1) Higher H-Vote for Review 1 and no H-Vote on Review 2 (N=37)
(2) No H-Vote for Review 1 and higher H-Vote for Review 2 (N=33)
(3) Low H-Vote for Reviews 1 and 2 (N=38)

The profiles of these adult participants are diverse: male (66.7%), female (33.3%); age 20-29 (54.6%), 30-39 (26.9%), 40-49 (13%) and 50-59 (0.9%).

To test the hypotheses, we used binomial tests since the dependent variable is binary (Yes or No) and the sample sizes were small. To evaluate these sub-dimensions of the H-Vote (Learning and Decision), we asked respondents to what degree (with a 5-point Likert-scale from "strongly disagree" to "strongly agree") they agree with the statement: *Above review is helpful to me to learn (or to make a purchase decision) about this product*. In other words, we refined the binary (yes or no) construct into the 5-point Likert-scale. This enabled us to use ANOVA (having those

[3] This was as of late April, 2011.

sub-dimensions as dependent variables) to explore the possible impact of consumer attributes such as gender, age, ethnicity and income.

4. Results

For the favorable review, the respondents rated 60.0%, 57.9% and 54.4% on H-Vote for high, no, and low H-Vote, respectively. The resulting numbers on the H-Vote rate are obviously similar and not statistically different from each other. In other words, different H-Votes did not affect the vote of survey respondents. Therefore, H1 is supported.

In contrast, the survey respondents voted differently on the critical (unfavorable) reviews for each of these three treatment conditions. When a high H-Vote was shown, the respondents voted lowest (73.7%). When the low H-Vote was displayed, the same review received the highest H-Vote at 87.5 %. This is a significant, 13.8% increase (*p* = .016) from the lowest H-Vote of 73.7%. When no H-Vote was shown, the respondent rated somewhere in between. Such a pattern supports H2.

	H-Vote Cue Shown vs. Resulting Average H-Vote by Respondents	
Review Valence	Favorable (Positive)	Critical (Negative)
High H-Vote Cue	100% shown → 60.0% (-40%, *p* = .000)	79% shown → 73.7% (-5.3%, *n.s.*)
No H-Vote Cue	Rating Not Shown → 57.9%	Rating Not Shown → 78.8%
Low H-Vote Cue	46% shown → 54.4% (+8.4%, *n.s.*)	45% shown → 87.5% (+42.5%, *p* = .000)

Table 1: Summary of H-Votes under Three Different Treatment Conditions

Measuring the review helpfulness by using the Likert-scale for Learning and Decision, we can see the overall patterns of results in line with H1 and H2. At the same time, the two constructs – Learning and Decision – reveal their differences.

The rating numbers in Table 2 show more variations for the critical (negative) review than for the favorable (positive) review. The results of ANOVA with Learning as the dependent variable had no significant factor/covariate, regardless of the valence of the review. That is, the Learning helpfulness rating was fairly stable whether a high, low or no H-Vote was displayed, taking into account any combination of different consumer attributes (e.g., gender, age, ethnicity and income).

In contrast, the Decision rating clearly showed more sensitivity to which H-Vote was shown and to consumer attributes than the Learning rating. For the favorable (positive) review, the Decision rating had two significant factors: age (*p* = .080) and income (*p* = .005) with F = 4.26 (*df* = 2, *p* =. 017). In other words, the Decision rating was higher as the age of consumers was *higher* (β = .184), but the income level of consumers was *lower* (β = -.297). The Decision rating of the critical (negative) review had two significant factors: income (*p* = .070) and treatment (*p* = .096) with F = 2.51 (*df* = 3, *p* =. 063). That is, the Decision rating for the negative review was higher as the displayed H-Vote was higher (β = .183), and as the income level of consumers was lower (β = -.205). H1 and H2 are further confirmed by the fact that the displayed H-Vote affected the Decision rating only for the critical (negative) review. In addition, the Decision rating is sensitive to income. Such is not the case for the Learning rating.

Review Valence	Favorable (Positive)		Critical (Negative)	
Sub-Dimensions	Learning	Decision	Learning	Decision
High H-Vote Cue	3.03 (1.18)	2.80 (1.18)	3.55 (0.86)	3.53 (0.92)
No H-Vote Cue	3.26 (0.79)	3.21 (0.96)	3.64 (1.27)	3.36 (1.27)
Low H-Vote Cue	3.00 (1.20)	2.79 (1.11)	3.97 (1.12)	3.91 (1.15)

[The number in each cell is the helpfulness rating mean: 1 is least helpful, and 5 is most helpful (for product learning or purchase decision-making. The numbers in parentheses are standard deviations.]

Table 2: Summary of Learning/Decision Helpfulness under Three Different Treatment Conditions

5. Implications and Conclusion

There are several important implications for practitioners and researchers. First, there are more positive reviewers than negative ones (Pinch & Kesler, 2011), and most shills are positively biased (Streitfeld, 2011). However, consumers seem to adapt to such a reality and are not affected by these displayed “false positive” H-Votes. Those vendors and retailers who use shills to write positive reviews with a high H-Vote might think their strategy will work. According to our research, consumers are not manipulated by such a strategy; consumers are smarter than vendors and retailers are hoping for.

Second, if a vendor wants to “torpedo” critical reviews on its products, should it hire shills to vote them down from the prominent position by manipulating the H-Vote? Our study shows that such a manipulation actually makes consumers feel those critical reviews are more helpful than they really are. The reason behind this is probably the quality of critical reviews exceeding the expectation of consumers based on the manipulated H-Vote. Thus, this type of manipulation is contraindicated for the vendor.

Third, a theoretical contribution of this exploratory study is to show the relation between the Guttman-scale H-Vote and the Likert-scale sub-dimensions (Learning and Decision) of the H-Vote. The Learning and Decision helpfulness ratings generally follow the ratings of the H-Vote. However, the Decision rating is more sensitive than the Learning rating regarding (a) how the review is presented and (b) who the consumer is. Semantically, the Learning and Decision helpfulness ratings are not identical. Consumers vote regardless of their goal orientations at Amazon.com. Therefore, the H-Votes are mixed bags, and may not be as useful as the Decision ratings if vendors and retailers want to identify the attributes of consumers who find certain reviews helpful or not.

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 sample. However, the common assumption that we, as consumers, vote the same way regardless of displayed helpfulness votes is not the case, at least for two bestselling vitamin products (credence goods). Future studies can investigate the validity of this finding using other types of common products.

References

Bhattacherjee, A. (2001). Understanding Information Systems Continuance: An Expectation-Confirmation Model. *MIS Quarterly, 25*(3), 351-370.

Darby, M. R., & Karni, E. (1973). Free Competition and the Optimal Amount of Fraud. *Journal of Law and Economics, 16*(1), 67-88.

Hu, N., Liu, L., & Sambamurthy, V. (2011). Fraud detection in online consumer reviews. *Decision Support Systems, 50*(3), 614-626.

Kahneman, D. (2003). A perspective on judgment and choice: Mapping bounded rationality. *American Psychologist, 58*(9), 697-720.

Moyer, M. (2010). Manipulation of the Crowd. *Scientific American Magazine, 303*(1), 26-28.

Mudambi, S., & Schuff, D. (2010). What Makes A Helpful Online Review? A Study of Customer Reviews on Amazon.com. *MIS Quarterly, 34*(1), 185-200.

Oliver, R. L. (1977). Effect of Expectation and Discontinuation on Postexposure Product Evaluations: An Alternative Interpretation. *Journal of Applied Psychology, 62*(4), 480-486.

Pinch, T., & Kesler, F. (2011). *How Aunt Ammy Gets Her Free Lunch: A Study of the Top-Thousand Customer Reviewers at Amazon.com*.Unpublished manuscript.

PowerReviews. (2011) The 2011 Social Shopping Study. Available at http://www.powerreviews.com.

Sigelman, L., & Knight, K. (1983). Why Does Presidential Popularity Decline? A Test of the Expectation/Disillusion Theory. *The Public Opinion Quarterly, 47*(3), 310-324.

Smith, J. S. (1960). The Use of the Guttman Scale in Market Research. *The Incorporated Statistician, 10*(1), 15-28.

Streitfeld, D. (2011). In a Race to Out-Rave, 5-Star Web Reviews Go for $5. *New York Times*. Retrieved from http://www.nytimes.com/2011/08/20/technology/finding-fake-reviews-online.html?_r=2&scp=8&sq=retail&st=nyt

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science, 211*(4481), 453.

Weiss, A. M., Lurie, N. H., & MacInnis, D. J. (2008). Listening to Strangers: Whose Responses Are Valuable, How Valuable Are They, and Why? *Journal of Marketing Research, 45*(4), 425-436.

Zhu, F., & Zhang, X. (2010). Impact of Online Consumer Reviews on Sales: The Moderating Role of Product and Consumer Characteristics. *Journal of Marketing, 74*(2), 133-148.