Do People Really Experience Information Overload While Reading Online Reviews?

Bum Chul Kwon, Sung-Hee Kim, Timothy Duket, Adrián Catalán, and Ji Soo Yi

July 2, 2015

Abstract

Online consumer reviews have become a substantial component of e-commerce and provide online shoppers with abundant information about products. However, previous studies provided mixed results about whether consumers experience information overload from such a vast volume of reviews. Thus, we investigate how users perceive products depending on various numbers of reviews (from 0 to 3,000 reviews) and different review valences (generally positive, generally negative, and divided). We conducted two crowdsourced studies with 1,783 participants. We found no clear evidence to suggest that information overload increases as the number of reviews increases. Instead, our participants relied on a very limited number of reviews in making purchase decisions. In addition, we observed that the review valence affected how the participants used different information sources from the interface. Based on our results, we provide a set of interesting implications and design guidelines.

Keywords: Online reviews, information overload, review valence, crowdsourced study

1 Introduction

Online reviews, a type of electronic word-of-mouth (eWOM), have impacts on individuals’ purchasing behavior because people desire to refer to others’ opinions and experiences (Hao, Ye, Li, & Cheng, 2010; Y. Liu, 2006). Online reviews are widely and vigorously adopted, so it is relatively easy to find products on popular e-commerce websites that have thousands of reviews. However, such a high volume of online reviews appears to provide not only benefits but also challenges (Etzion & Awad, 2007). Despite valuable information in consumer reviews, reading thousands of reviews is virtually impossible. Thus, a plethora of research studies (e.g., Carenini, Murray, & Ng, 2011; B. Liu, 2010) have been conducted to help online consumers make sense of large quantities of online reviews.

Prior research studies have shown mixed results on the effect of the volume of reviews on consumer's purchasing intention. Park et al. (2006) reported that consumers experienced information overload and lowered their purchasing intention when there were more online reviews. On the other hand, some other studies reported that the number of reviews is positively correlated with product sales and consumers' confidence and trust (e.g., Duan, Gu, & Whinston, 2008; Y. Liu, 2006). In addition to the conflicting results on the review volume, there are other factors beyond the number of reviews influencing online consumers' perceptions and behaviors (purchasing), and one of the salient factors reported is review valence (whether reviews are positive or negative) (Chevalier & Mayzlin, 2003; Hao et al., 2010). Thus far, it has not been established how the volume and the valence of consumer reviews affect consumer's information overload.

Therefore, the goal of this paper is to understand how consumers perceive information overload when they encounter varying numbers of reviews and different kinds of review valence while purchasing goods online. We organized our research questions in the following way:

R1 Will online consumers experience more information overload as the number of online reviews increases?

R2 Will online consumers experience different levels of information overload as the review valence changes? How will the number of online reviews and the review valence affect online consumers experience?

In order to answer these questions, two crowdsourced studies were conducted using Amazon Mechanical Turk (MTurk). In Study 1, we varied the number of online reviews and investigated whether the changes influenced research participants' perception followed by a small interview study with research participants. In Study 2, on top of varying the number of reviews, we also changed the valence of online reviews to investigate the influences of review valence.

We believe that this study has the following contributions:

- The results of this study help us question key assumptions in relevant studies (i.e., a high volume reviews causes information overload).

- This study provides evidence-based design guidelines for presenting online reviews.
- The study demonstrates that a crowdsourced approach is a viable methodology to strike a balance between internal validity and external validity for large-scale studies with human participants.

2 Related Work

2.1 Review Volumes and Information Overload

These days, it is common for a product listed on an e-commerce website to have thousands of reviews. One of the most extreme examples is an electronic reading device, called "Kindle Fire (the first generation)," which has 23,100 reviews on Amazon.com as of October 10, 2014. Other studies also report similarly massive volumes of online reviews. Chen et al. (2004) reported that 25% of books used in their study had between 73 and 4,607 reviews. Khare et al. (2011) conducted a study using control sizes of either 62 or 3,470 reviews on movies on the Yahoo! Movies website.

Prior research studies investigate how the increasing volume of reviews affects the purchase decision making process. Such research studies on the effects of online review quantity often use one of two approaches; (1) controlled studies with human participants and (2) studies analyzing publicly available data collected from online shopping websites, forums, and discussion boards.

The first group tests the effects of review volume on information overload by conducting human subject studies, in which participants are asked to make purchase decisions on online shopping websites. Park et al., (Park et al., 2006) tested the effects of the number of reviews (i.e., 1, 6, 12) by showing varying numbers of reviews without any other summary information. Following studies also tested different review volumes: 4 and 8 (Park & Kim, 2008), and 3, 9 and 27 (Park & Lee, 2008). These studies reveal that participants perceive higher information overload as the number of reviews increases. However, the largest number of reviews tested is 27, so the trend found in these studies may not hold true when consumers encounter much larger volumes of reviews. Furthermore, the four studies above did not provide some basic summary statistics, such as five-star ratings and their overall distributions which are common in leading online websites (e.g., Amazon.com). In addition, they controlled the review length to be three lines with a font size of 10, which does not reflect the realistic online reviews.

The second group infers consumers' perception of online reviews by investigating the correlation between review volumes and product sales. Such studies tend to include larger numbers of online reviews. For example, prior studies used 303.4 reviews per product (Y. Liu, 2006) and 1,350.24 reviews per product (Duan et al., 2008), in average, that are collected from e-commerce websites and compared with other performance metrics (e.g., sales or revenues). These studies argue that the increasing review volume leads consumers to have more confidence and trust in the product and is positively related to increasing product sales or revenue (Clemons, Gao, & Hitt, 2006; Dewan & Ramaprasad, 2007). Though obtaining and analyzing data from e-commerce websites increases the external validity of these studies, the correlation between review volumes and product sales does not guarantee causality. The volume of reviews could be a precursor (i.e., driving force), an outcome, or both of product sales (Duan et al., 2008). Thus, prior studies cannot show the effect of the review volume on individual consumer's perceived information overload.

In summary, we believe that there is a gap between these two groups of studies. Each group either sacrifices external validity (too small number of reviews and unrealistic interfaces in the first group) or internal validity (no control over different conditions to verify the causality). In addition, this study did not consider other factors (e.g., valence) that may influence the review reading behavior. This motivates us to conduct another study to bridge the gap between the two groups.

2.2 Review Valence

Through the extensive literature review, we also identified many other factors influencing online consumers' purchasing intention and information overload, such as valence of reviews, types of products (experience goods versus search goods) (Hao et al., 2010), types of reviews (attribute-value vs. simple recommendations) (Park & Kim, 2008), involvement of consumers (high involvement vs. low involvement) (Park et al., 2006), the length of reviews (Chevalier & Mayzlin, 2003), and the credibility of each review (Metzger, 2007). A search good is a product or service that has characteristics that are easy to evaluate and substitute; in contrast, an experience good (e.g., jewelry) is a product or service that has characteristics that are hard to evaluate and substitute (Nelson, 1970). Consumers' involvement can be categorized into two concepts: i) enduring involvement: consumers' genuine interest in a target product or service over a long period of time; ii) situational involvement: consumers' interest elicited in a particular situation. Credibility of reviews means whether the reviews are current, accurate, factual, and complete.

Out of many factors, we found that review valence, how positively or negatively a review or group of reviews is perceived by the reader, is one of most salient factors that may influence purchasing decisions and information overload. Hao et al. (2010) claims that reviews with negative valence have a greater impact on consumers when deciding whether or not to purchase a product. Consumers engage more while reading negative reviews (Sen & Lerman, 2007) and if the review fits their processing strategy, the influence of valence is greater (Park & Kim, 2008). On the other hand, Huang & Yen (2013) found that positive reviews are regarde more helpful to consumers. Though many studies reveal the impact of valence on consumers' purchasing decisions, it is unclear how the valence can influence the perception of information overload in the decision making process.

2.3 How do Consumers Deal with Information Overload?

Little research comprehensively surveys what kinds of strategies individual consumers employ to deal with information overload. However, from some studies, we could infer two general strategies that online consumers might use.

First, users might select and read a limited subset of the available reviews that fit their needs. According to a survey (Anderson, 2012), 67% of consumers reported that they feel they can trust a local business after reading 2 to 10 online reviews and ratings (25% said 2 to 3, 22% said 4 to 6, and 20% said 7 to 10). This suggests that people may selectively read only a small number of online reviews regardless of the number of reviews available. Badke (2010) claims that consumers may be more content with reading a subset of reviews in a short time rather than sifting through all of the reviews. When enough reviews are available, a consumer may select a handful of reviews that seem trustworthy and ignore the rest, even if the heuristic leaves thousands of reviews unconsidered.

Second, consumers may consider only simple summary statistics, such as average rating or distribution of consumer ratings (e.g., 5-star ratings). With such averages and distributions, the overall valence reviews may become clear to the reader at a quick glance. Consumers tend to reject products with low average ratings when the products are not familiar to them (Ozakca & Lim, 2006). Lelis and Howes (2011) also argue that consumers tend to rule out individual reviews with low ratings so that they can decrease their information processing time and effort. To the best of knowledge, no study reports individual consumers' review reading behaviors and strategies through human subject experiments. Thus, we also decided to observe what kinds of strategies individual consumers employ in order to retrieve information from many reviews with a qualitative survey and an interview.

3 Study 1: Information Overload and Review Volume

The goal of this study is to answer our first research question (R1). We tested twelve levels of volume which range from 0 to 3,000 reviews. The twelve levels were selected to cover the range of the review numbers that were tested in two groups of prior studies discussed in the Background section. Since the volume of reviews is the main factor of this study, other factors that might influence the consumers' perception were either controlled or randomized.

3.1 Methods

3.1.1 Participants

A total of 999 participants (age range: 18 to 78 with an average of 28.9 ($\sigma = 9.26$); 252 females) were initially recruited through Amazon Mechanical Turk (MTurk). crowdsourced studies are subject to having unwanted participants who are only interested in earning money and do not participate in the studies faithfully under proper guidance (Paolacci, Chandler, & Ipeirotis, 2010). Thus, we created a simple survey question in order to filter out such unwanted participants as other researchers do (e.g., Shaw, Horton, & Chen, 2011). The survey question asked,"What type of product did you review?" (answer: a GPS navigation device). Participants who failed to answer it correctly (336 participants) as well as ones who did not complete the survey (19 participants) were excluded from data analysis. The data from the remaining 614 participants (241 females) were analyzed and reported here. The number of participants for each level of treatment ranged from 48 to 54. The spread of education levels of the participants was as follows: 4-year University or College: 50.3%; Graduate school or above: 31%; and High School: 17%. 96.5% of participants reported that they read online reviews when making purchases. After a pilot study, we decided to offer \$0.10 as compensation for participation to this experiment. It took ten days for all of participants to complete the experiment.

3.1.2 Experimental Design

[Table 1 about here.]

This study was designed as a between-subject study with one factor (the number of reviews) with 12 levels (0, 1+, 1-, 6+, 6-, 12, 30, 50, 100, 300, 1,000, and 3,000) as Table 1. We did not use a within-subject design because participants can learn about products and reviews through it. Factors other than the number of reviews were strictly controlled. Since this study used consumer reviews collected from Amazon.com, consumer ratings were recorded the five-star rating system, in which 1-star indicates the most negative review and 5-star indicates the most positive review. Thus, we used "star rating" as consumer rating in this study for convenience. To control the overall valence of review, we selected reviews randomly and proportionally (5-star: 50%; 4-star: 26%; 3-star: 8%; 2-star: 7%; 1-star: 10%), so that the distribution of ratings in each matches with the original distribution of ratings, which we call, "F-distribution," because the shape of the distribution resembles the shape of the letter F as shown in Figure 1. However, when the total number of reviews was small (i.e., 1 and 6), maintaining the F-distribution became impossible. Thus, we divided these two conditions into positive cases (i.e., 1+ and 6+) and negative cases (i.e., 1- and 6-) as shown in Table 1. In addition, we included the level of no reviews to clearly examine the effect of no consumer reviews.

[Figure 1 about here.]

We controlled the involvement level by introducing a scenario, which will be explained in the following section. In order to maintain a realistic environment, we randomized following factors: review type (attribute-value versus simple recommendation), review length, review credibility, review order, and participant gender.

3.1.3 Procedure

Our study followed Hwang's Internet study (2014) that starts with online shopping and ends with survey. The difference from Hwang's study is that we developed our own website to track users' interactions. We posted our task on the MTurk platform. In the post, a simple description of this experiment was written. If the participants agreed to participate after reading the description, they were redirected to our separate experimental website. The following task scenario was first given to maintain a high level of involvement from participants. We asked participants to evaluate the given GPS device for an imaginary person, named "Vina":

> Vina is a 35-year-old housewife who lives with her husband and two children. The family recently purchased a house in a new, unfamiliar city. Vina will be driving her children to and from school and other various activities for the next several years before they can drive themselves. The family also hopes to take advantage of their new location and take weekend trips to new parts of the city and surrounding areas. For these reasons, Vina is researching the available GPS devices currently on the market and hopes to find an affordable product that will be reliable both in and out of the city.

By introducing this scenario, we intended to block the effects of individual needs, preferences, and financial status and to encourage participants to faithfully read reviews and provide the best recommendation. After reading this scenario, participants entered the interface of online shopping website, in which they read description and consumer reviews about a GPS product. Whenever they made their purchase decision, they were asked to click on "Continue to Survey." Then, a post experiment survey was given. After they filled out survey completely, the website provided an alphanumerical code that can be used to claim compensation on MTurk.

3.1.4 Measures

Four different types of measures were used to analyze a consumer's perception of reviews. These include consumer purchasing intention (Cheung & Thadani, 2010), information overload (Park et al., 2006), and perceived quantity and quality of reviews (Lee, 2009). We adopted and modified ten survey statements to quantify the four measures as shown in Table 2. The original statements were captured from Park et al. (2006) and Lee (2009). We modified the statements for this study. All statements were converted from questions to descriptive sentences. This conversion can make all scale values written in the same scale range from 'Strongly Disagree (1)' to 'Strongly Agree (7)' including 'Neither Agree nor Disagree (4).' We did this to ensure participants not to be confused from different scale ranges (e.g., More Satisfied, Highly Likely) for different questions. We used one statement for Purchasing Intention because participants were given a recommendation task. Including his or her choice could also introduce external factors like personal preferences. We paraphrased original statements for Information Overload. We changed a word "suffer" of the fourth statement to "experience" because "suffer" could be perceive much intensive experience. We paraphrased all statements for Perceived Quality and did not change wordings. There is no original statement available for Perceived Quantity. Lee (2009) asked participants to recall whether the number of reviews they were given is larger than the number of reviews for other irrelevant products. However, we did not want to measure whether participants remember the approximate number of reviews in their given settings. We believed that participants will remember the approximate number due to the explicit number presented on the experimental website. Rather, we wanted to measure whether participants perceive the given number of reviews as large or small. Thus, we created the two statements in Table 2 to ask their perception on the review volume.

In addition, we asked participants to report how each of six different sources of information (product title, product price, product description, consumer reviews, average rating, and distribution of ratings) affected their final recommendation in percentage; the six added up to 100%. We also collected demographic information such as age, gender, education, and online shopping experience. In addition, we captured how the participants navigated through the online reviews by three interactions: page navigation (clicking to the next page of reviews), sorting by date or by helpfulness, and filtering by star-rating.

[Table 2 about here.]

3.1.5 Dataset

To conduct this experiment, product description and consumer reviews are necessary. In order to create a realistic e-commerce setting, we decided to use a commercially available product and its consumer reviews that are listed in online shopping websites. At the same time, we decided to mask the product names and its subcomponent names to block the effect of brand recognition. The following steps were used to gather necessary data for this experiment:

S1 We comprehensively reviewed products that are listed as best sellers of Amazon.com (`http://www.amazon.com/Best-Sellers/`) as of June 22, 2012.

S2 Among 2,791 products found in Step 1, we selected a single product ("Garmin nüvi 1490LMT 5-Inch Bluetooth Portable GPS Navigator with Lifetime Map & Traffic Updates," `http://www.amazon.com/Garmin`

-Bluetooth-Portable-Navigator-Lifetime) based on the following criteria: 1) the product should have more than 3,000 reviews to create our treatment levels, 2) the distribution of ratings should be in an F-distribution (more positive reviews than negative reviews), which is more commonly found among products with the high review volume, 3) the product should be a search good because the perception of the product should be less influenced by personal preference or prior experiences rather than experienced goods, 4) the product should be attractive to the largest possible population, and 5) the product should contain some technical specifications so that participants are encouraged to study them.

S3 We scrapped product information (e.g., name, description, and price) and review information (e.g., review author's name, star rating, the time of review submission, and review helpfulness vote). We used only the first paragraph of the product description and removed any additional advertisements in order to reduce any distraction while still providing succinct and necessary background information. Qualification about the reviewers specific to Amazon.com (e.g., "#1 reviewer" and "Vine$^{\text{TM}}$voice") were also removed.

S4 We anonymized review writers and masked the product name and its product specific information (e.g., "Garmin nüvi") with arbitrary names (e.g, "Cartographer Guide") to avoid the influence from the brand name.

The selected product had 3,088 reviews and the following proportions of star ratings: 5-star: 50%; 4-star: 26%; 3-star: 8%; 2-star: 7%; and 1-star: 10%.

3.1.6 Design of Experimental Website

The interface of the experimental website was designed to mimic the basic layout and core features of Amazon.com, which is arguably one of the most popular e-commerce websites, for a realistic and familiar online shopping environment. In addition to the interface, we also implemented three key interactions (i.e., page navigation, filter by star rating, and sort by date and helpfulness) used in Amazon.com. Figures 2 and 3 show screenshots of the interface with one review and with 3000 reviews, respectively. The top portion of the page shows product title and product description. We removed the product image because it may bias participant's responses. Below product description, the website shows the distribution of star ratings, the number of reviews, and the average star rating. The star rating distribution is shown in a vertical histogram as on Amazon.com, and clicking on each bar displays all comments with the corresponding rating. Hovering over each histogram bar triggers a tooltip showing the percentage of reviews with the star-rating (e.g., 50% for five-star rating). At the bottom, reviews are presented in a list. The review list can be sorted by review helpfulness and review date. The website displays 10 reviews per page initially sorted by review date. Each review comprises its star rating, review title, review date, review text, and review helpfulness. While closely following the interface of Amazon.com, we eliminated other commercial features such as advertisements and promotions to minimize external influences.

[Figure 2 about here.]

[Figure 3 about here.]

3.2 Results

3.2.1 Manipulation and Control Checks

In this section, we test the validity of our experimental setting using two measures: perception of quantity and quality of reviews. We follow the manipulation check method utilized by previous research (e.g., Hao et al., 2010; Park et al., 2006). If the participants performed the experiment properly, the perception of quantity should increase as the number of reviews increase while the quality should not be different across twelve treatment levels. Mean values of two measures of the perception of the review quantity were computed (see Table 2) and the ANOVA analysis indicated that the number of reviews turned out to be a significant factor ($F(11,602) = 16.84$; $p < 0.01$). As the number of reviews increased, the mean value also increased (see Figure 4(a)). Mean values of three measures of the perception of the review quality were also calculated and the ANOVA analysis indicated that the number of reviews turned out to be significant ($F(11,602) = 8.80$; $p < 0.01$). We used mean values by following prior research studies (Park et al., 2006; Lee, 2009) that we derived and modified survey statements from. The Tukey test reveals that the zero review condition is different from the rest of eleven conditions (for all pairs, $p < 0.05$), but there is no difference in other pairs of conditions (see Figure 4(b)). This demonstrates the validity of the results and that all participants performed the task and completed the survey properly.

[Figure 4 about here.]

3.2.2 Purchasing Intention

Since purchasing intention does not satisfy the assumption of the ANOVA test, the Kruskal-Wallis test was employed. It revealed a significant effect of the number of reviews on purchasing intention ($\chi^2(11) = 78.98$; $p < 0.01$). The Pairwise Wilcoxon Rank Sum test was conducted as a post-hoc test with Bonferroni correction to adjust errors. Only a few pairs of conditions were identified as significantly different. The purchasing intention with volume '1-' was lower than the rest of conditions (for all cases $p < 0.01$). When there were no reviews, the purchasing intention was lower than when the volume was 1 ($p = 0.01$) and 3,000 ($p = 0.02$). Other than the cases above, the mean values were above 4.9 (5 corresponds to "Somewhat Agree" on Likert scale), showing the number of reviews did not affect purchasing intention.

3.2.3 Contribution of Six Information Sources

The contribution levels of the six information sources (product title, product price, product description, consumer reviews, average rating, and distribution of rating) were initially collected as percentages (e.g., "23% of my decision is based on the product title.") which sum to 100%. However, since the contributions of different information sources are not independent, parametric statistical tests cannot be used, so we convert the percentage metrics into rankings of contributions (contribution level, henceforth) from 1 (the lowest contribution) to 6 (the highest contribution). Then we employed a Kruskal-Wallis test to observe the influence of the number of reviews on levels of contributions as shown in Table 3. In every case except for the distribution of ratings, the test revealed a significant effect of the number of reviews on the contribution level.

[Table 3 about here.]

Figure 5 shows the contribution level of each information source separately. The first group (product title, product price, and product description) presents information determined by product sellers. The second group (consumer reviews, average rating, and distribution of rating) shows information provided by consumers. A different direction of trend from the two different information sources: the product seller and consumers. As the number of reviews increases, the contribution level of consumer reviews and average rating increases and those of price and product description decrease while the contributions from the distribution of ratings do not change much.

[Figure 5 about here.]

3.2.4 Information Overload

[Figure 6 about here.]

We use mean values of four measures (i.e., perceived information overload, satisfaction of decision, confidence of decision, and confusion during the task) as the measure of information overload. An one-way ANOVA was conducted to compare the effect of the number of reviews on information overload. There was no significant effect of the number of reviews ($F(11,602) = 1.35$; $p = 0.19$) as shown in Figure 6.

3.2.5 Interaction Results

From the interaction logs recorded from the interface, we can see that a majority of participants did not use the available navigation techniques. The most frequently used interaction was page navigation which was used by 64 participants (10.4%). Sorting by date and helpfulness of the reviews was used by 48 participants (7.81%). Finally, filter by star-rating was used by 21 participants (3.42%). This indicates that the participants who did not apply additional interactions made their decision using only the first given page of reviews. Table 4 shows the number of interactions performed by participants in twelve different review conditions.

[Table 4 about here.]

3.2.6 Interview Results

An interview study was conducted to investigate the process participants took while performing our tasks in Study 1. In the survey from Study 1, we asked participants whether they are interested in joining an interview via a tele-conferencing software (Skype). After we collected data for Study 1, we sent out emails to invite participants for the interview study. As a result, we recruited six participants (2 females) who participated in Study 1, who will be referred as P1 to P6, henceforth. P1 to P6 had 1,000, 50, 30, 1,000, 1,000, and 50 reviews respectively during their trials. This interview study was conducted less than three days after their trials. Though the interview focused on revealing their task process and strategy, we let the conversation naturally extend to general online shopping experiences. Throughout interview sessions, we could observe three interesting user behaviors.

Cutting the number of reviews to read. Participants did not read all of reviews. Instead, they tended to make recommendations based on ten reviews in the first few pages. All participants confirmed that they read only up to four or five pages at maximum when they do online shopping.

Participants followed some strategies to cut the number of reviews they need to read. P1, P5, and P6 read negative reviews first and decide whether or not they need to read more reviews. P6 revealed that he only reads ten reviews showing on the first page because it provides enough information. P1 revealed her strategy vividly using her most recent experience with purchasing a can opener in an online shopping website:

> P1: "I first sort products by price because price is the most important thing I care about. Then, I enter the page of the cheapest product, then start reading negative reviews. Reviews were saying that a teeth of the can opener tends to break really fast. Then, I don't even bother with positive reviews [of the can opener] at that point."

Another strategy employed by P2 and P3 was to use the average rating to cut the number of reviews. Both of them had a threshold which they brought up throughout their online shopping experiences. P2 reported that she only considers products that have more than three quarters of four- and five-star ratings out of all reviews. P3 reported that he usually filters out any products that have an average rating less than 4.0 out of 5.0 given that there are more than 5 reviews in total:

> P3: "I first check what average star ratings the product has. If the product is 4.0 out of 5.0, then I go onto the next step. But, if the 4.0 out of 5.0 rating is out of less than five reviews, I wouldn't even consider the product because it's not enough reviews."

Judging the characteristics of negative reviews. All interview participants revealed that they regard negative reviews very importantly because they tend to include defects or errors that can come out of a long-term use. P1, P2, P3, and P4 reported common strategies to read negative strategies. They revealed that negative reviews should be read more carefully because the reviews often contain user's faults or complaints. In order to filter those kinds of personal issues, they usually try to ensure those issues are confirmed from other reviews as well.

> P4: "[...] I also scan these disadvantages [negative reviews] and see they are from products or consumers' preferences. Sometimes people don't like a product, and they try to report that online. So, I try to avoid falling into such things. I also try to take a look if other consumers' reviews also support such issues. "

Searching attribute-based reviews. When participants investigate reviews, they tend to scan reviews and find ones that describe individual product features and their advantages and disadvantages. In particular, they reported that they prefer reviews written in an organized fashion, such as using a table or bullet points. P4, P5, and P6 reported such behaviors, and P5's quote shows his desire to have some organization in text.

> P5: "I wanted more organization especially in terms of pros and cons. I feel it would be much useful if pros and cons for each specification are organized in some reviews. Unfortunately, I could not find such kinds of reviews. Text is too descriptive. Only if I have time, I would read all."

3.3 Discussion

Surprisingly, we did not find statistical significance to suggest that the number of reviews affects information overload. To confirm this finding, we compare the levels of perceived quantity of reviews with information overload (see Figure 7). The blue plots are from the perception of quantity and the red plots are results from information overload. It is clear that the mean perception of quantity increases but the information overload neither increases nor decreases as the number of reviews increases. This implies that our study participants were clearly aware of the different numbers of reviews, but they did not experience more information overload. This result conflicts with previous research (e.g., Park et al., 2006) in which the large number of online consumer reviews caused information overload. We believe that there are several reasons why our results did not replicate the results of Park et al. (2006) in our study. One difference between this study and Park et al. (2006) is that the experiment website in this study shows summary statistics such as average rating and distribution of ratings as shown in Figure 2 while Park et al.'s study (2006) did not show any summary statistics to their participants.

[Figure 7 about here.]

These findings are in line with the previous discussion of the interview results. Although participants are aware of the varying number of reviews, we have shown that they selectively read a limited number to alleviate their own information overload.

4 Study 2: Information Overload, Review Volume, and Review Valence

Though Study 1 answered our first research question (R1), the results are not fully generalizable because the online reviews in Study 1 are heavily positive (F-distribution). It is unclear whether the trends will stay the same even when the general valence of reviews is mixed (C-distribution) or negative (L-distribution). Thus, Study 2 was designed to elicit the effects of the different distributions of star ratings on consumer's perception of reviews.

4.1 Methods

4.1.1 Participants

We also conducted Study 2 using MTurk. As we did in Study 1, we offered \$0.10 as compensation for completion of the experiment. A total of 784 participants (age range: 18 to 67 with an average of 28.52 ($\sigma = 9.12$); 310 female participants) were initially recruited from MTurk. Applying the same filtering method as in Study 1, the data of 171 participants were excluded from analysis. Additional data of 64 participants were excluded due to incomplete survey submission. Their education levels were as follows: 4-year University or College: 40.1%; Graduate school or above: 26.9%; and High School: 26.7%. 95.9% of participants reported that they read online reviews when making a purchase. We collected data for 26 days to recruit these participants, not including the period of time spent collecting data for Study 1. For further analysis in Study 2, remaining data from 549 participants were used (275 for C-distribution; 274 for L-distribution) for analysis on top of data from Study 1. The number of participants for each level of treatment ranged from 40 to 61.

4.1.2 Experimental Design

Study 2 also adopted the between-subject design as in Study 1 so that it can block the learning effects and fatigue effects from subsequent trials. We also blocked participants from Study 1 to enter Study 2 using their MTurk ID. In Study 2, review volume and the distribution of star-ratings were varied as follows.

We newly introduced two types of distributions: C-distribution (heavily conflicting positive and negative reviews) and L-distribution (heavily negative reviews) as Figure 8 shows. The volume of reviews were varied in the range of 6, 12, 30, 50, 100, and 300. Other levels such as 0, 1-, 1+, 1,000, and 3,000 were excluded from Study 1 because the two lower levels, 0, 1-, and 1+, had no particular distributions and 1,000 and 3,000 reviews were not feasible because our product does not have enough 1-star and 2-star reviews for the distributions. Removing these levels was necessary in order to use the same GPS product to be able to compare across different distributions. We could have used another product for Study 2, which has more negative reviews, but we did not do in order to avoid any confounding factors. Other factors were controlled or randomized as was done in Study 1. We also provided the same survey questionnaire.

[Figure 8 about here.]

4.2 Results

4.2.1 Purchasing Intention

The Kruskal-Wallis test revealed a significant effect of distribution on purchasing intention ($\chi^2(2) = 301.81$, $p < 0.01$). All distributions were shown to be significantly different ($p < 0.01$) from each other via a post-hoc test with Bonferroni correction. Figure 9 shows the actual number of participants grouped by their rating for recommending the product. For F- and L-distributions, the overall purchasing intention follows the valence of reviews. When there are conflicting reviews as in C-distribution, the recommendation also split into both directions. Therefore, this result also suggests that manipulation on this experiment was successful.

[Figure 9 about here.]

4.2.2 Contribution of Information Sources

[Table 5 about here.]

After we converted the percentage of contribution into the contribution level as in Study 1, we applied the Kruskal-Wallis test. The test revealed a significant effect of the distribution of star-ratings on all of the six information sources except for product title as Table 5 shows. Figure 10 shows that the contribution level of product price was higher in the F-distribution than C- and L-distributions. The contribution level of product description in the F-distribution was higher than in the C- and L-distributions and that in the C-distribution was higher than in the L-distribution. The contribution levels of consumer reviews and their average rating in the L-distribution were higher than in the F- and C-distributions. The contribution levels of distribution of rating in the C- and L-distributions were higher than in the F-distribution (Note: All of the pairwise comparison were done with Wilcoxon Signed-Rank Test using Bonferroni correction at the $\alpha = 0.05$ level of significance).

[Figure 10 about here.]

4.2.3 Information Overload

Information overload was calculated in the same way by averaging the four measures as in Study 1. Our initial approach was to use the number of reviews, distribution, and the interaction of the two as independent variables. However, the interaction effect did not come out to be significant ($F(10,843) = 0.77$, $p = 0.65$), so we removed it from the model. We applied a two-way ANOVA without interaction, and the distribution was significant ($F(2,843) = 8.57$; $p < 0.01$; mean values of 2.7, 2.6, and 2.4 for F-, C-, and L-distributions, respectively). There was no significant difference between the number of reviews ($F(5,843) = 0.42$, $p = 0.83$). The Tukey test revealed that information overload in the F-distribution was significantly larger than that in the L-distribution, but other pairs did not show statistically significant differences (see Figure 11).

[Figure 11 about here.]

4.2.4 Interaction Results

Similar to Study 1, participants did not perform interactions much. Across three distributions (F-, C-, and L-distributions), participants used page navigation the most (15.70%, 13.09%, and 16.42%, respectively), followed by using the sorting feature (7.30%. 5.45%, and 5.47%, respectively) and filtering by star-rating (2.56%, 2.90%, and 2.55%, respectively). Tables 6 and 7 show the number of interactions performed in two different distributions respectively across the twelve levels of review volume.

[Table 6 about here.]

[Table 7 about here.]

4.3 Discussion

Information from consumer reviews seems to affect consumer decision making differently according to distribution types. Participants reported that the contribution levels of the six information sources were different among the three distributions. The six sources of information can be categorized into two groups. In Figure 10, the two groups tend to show different trends. As Lellis and Howes (2011) suggests, participants may be more impacted by negative reviews. Thus, participants might have perceived the contribution levels of average rating and distribution higher in the C- and L-distributions which show a more prevalent existence of negative reviews.

Participants felt more information overload as the distribution changed to more positive reviews from L- to C-, and to the F-distribution. Although we can observe this increase in information overload as more positive reviews are included in the distributions, it is important to note that the mean value for information overload in each distribution was less than 3 on the 7-point Likert scale. This supports the claim that consumers are skilled at using their own strategies to control the vast number of reviews so that they reduce their information overload. We can also make an interesting observation when we plot information overload against each level of purchasing intention in Figure 12.

[Figure 12 about here.]

We found a sign that perceived information overload may be related to purchase intention. Figure 12 shows that the lowest information overload in each distribution corresponds to low purchasing intention (e.g., 1 or 2 on the Likert scale) and the highest information overload comes with mid to high purchasing intention (4 or 5 on the Likert scale). In a closer look, we can see the following scenario. With positive reviews (F-distribution), participants felt less information overload when they decided to recommend the product (e.g., 5, 6, 7 on the Likert scale). With mixed or negative reviews (C- or L-distributions), participants felt higher information overload when they decided to recommend the product (e.g., 5, 6, 7 on the Likert scale) than when they decided not to recommend the product (e.g., 1, 2, 3). We can conjecture that participants felt more information overload when they decide to recommend the product against the general consensus of consumer reviews. In such situations, they might be forced to read individual reviews more closely. This result calls for a future study on user interface designs and user interactions to support such behaviors that people tend to relieve their information overload differently against different review valence.

These observations also make sense with respect to the results from the six information sources. Recall that in Figure 10, the user-generated information sources (reviews, average rating, and distribution of ratings) were most heavily used in the C- and L-distributions. When the distribution of ratings and average rating clearly present controversial or negative opinions in the C- and L-distributions, people might find it easier to reject the product. On the other hand, participants who faced overall positive opinions in the F-distribution might have been more inclined to further inspect reviews and consider such seller-generated elements such as price more closely. In addition, people tend to seek more attribute-based reviews while making purchase decisions as revealed in Study 1. In other words, people seem to "reject" a product easily with other people's opinions but it takes more efforts to "purchase" a product. This behavior can also be explained by prospect theory (Kahneman & Tversky, 1979). Prospect theory suggests that while making a decision, loss is perceived as more significant than equivalent gain, which leads people to avoid losses as much as possible. This loss aversion could be a motivation for people to seek more information when they are recommending or purchasing a product.

The study results imply the reason that the results of this study conflict with Park et al. (2006). In their study, Park et al. (2006) did not show summary statistics of any form, which may have caused participants to feel more information overload with growing review volume. However, in our study, the existence of average rating and distribution of ratings may have lifted some of the burden from the participants, especially in the case of making rejecting decisions. We can conjecture that the summary statistics such as average rating and the distribution of ratings quickly served as a "scarlet letter" on the perception of the product so that they could easily reject the product.

The results also hint that we need better navigation tools to support consumers needs. Though participants revealed that they seek more attribute-based reviews in their interview, there is no feature to support such needs. Existing features like sorting and filtering were only lightly used by participants. Making purchase decisions with information from only the first ten reviews may not cause information overload but is likely suboptimal for the purchasing decision. Designers working within the human-computer interaction domain may need to create a visual navigation tool that allows participants to easily investigate the important content hidden in many reviews. One example is to implement a hierarchical sorting for product qualities (Cai & Xu, 2008), based on extracted information from consumer reviews.

On the other hand, we may conjecture that average rating and the distribution of ratings may yield rejections that come about too quickly. Even in the L-Shape distribution, we observe that some participants could still make purchasing decisions based on some of the reviews. Since the rating and distribution merely present sentiments in numbers, the information is rather incomplete. Designers may consider providing more descriptive summaries that can deliver more than numeric information.

5 Conclusions

We noticed that the number of online reviews does not directly contribute to information overload. However, this does not necessarily mean that large numbers of online reviews do not cause any issues. We found that participants in our study do not feel information overload not because they can make sense of all the reviews available, but because they largely ignore the existing data. Thus, existing technologies to summarize a large number of reviews still need to play a substantial role. Particularly, since we noted that many of our research participants rarely read beyond the first page, providing a useful first page is quite important.

Our study found that people feel less information overload as they encounter reviews that are negatively distributed. Negative reviews tend to play a more important role in influencing potential buyers. Negative reviews can contain flaws that can only be uncovered by consumers. As we found in our study, consumers want to rule out some products depending upon the existence of such issues. In the current interface, it is hard to collect such weaknesses easily from numerous consumer reviews. Such disadvantages as well as advantages of products need to be organized so that consumers can weigh their options more properly.

In addition, we also found that the crowdsourcing approach we used in Study 1 and Study 2 were a viable methodology to strike a balance between internal and external validity. Through the crowdsourcing platform, we could conduct a study with various research participants with a sufficient control, which could not be possible with either through controlled lab studies or using data from real environments.

5.1 Limitations and Future Work

There are several limitations to our study. First, we conducted crowdsourced studies, so there may be unfaithful participants even though every effort was made to filter them out. Second, prior knowledge on any GPS related products could affect the information process. Third, the results may not be directly generalized to the situation in which consumers purchase products for themselves. Since the experiment was designed to provide a purchasing situation for an unknown person, participants might not have put their best efforts in processing the information. Fourth, we did not include the product image, which might have affected participants' perception of the product. Fifth, some participants in our experiment might not fully engage in the task because they were from a crowdsourcing platform and a task was simple. These limitations need to be considered while readers receive insights from this manuscript.

Based on this preliminary study, more studies should be done in the following areas. We would like to test the effects of other factors on consumer decision making. Some of those factors are; 1) the type of products (experience goods and search goods), 2) the type of review (attribute-based and subject recommendation), and 3) credibility of reviews. In addition, we are designing a tool that consumers can use to make sense of reviews better. The tool will include information visualization of reviews in which consumers can navigate sections of reviews more easily.

6 Acknowledgements

We thank Nag Varun Chunduru and Zhihua Dong for their help in website design and literature review throughout this project. We also thank Sukwon Lee who provided constructive feedback on this study.

References

Anderson, M. (2012). Study: 72% of consumers trust online reviews as much as personal recommendations. *Search Engine Land*.

Badke, W. (2010). Information overload? maybe not. *Online 34.5: General Reference Center Gold*, *34*(5), 52–54.

Cai, S., & Xu, Y. C. (2008). Designing product lists for e-commerce: The effects of sorting on consumer decision making. *International Journal of Human-Computer Interaction*, *24*(7), 700–721.

Carenini, G., Murray, G., & Ng, R. (2011, June). Methods for mining and summarizing text conversations. *Synthesis Lectures on Data Management*, *3*(3), 1–130.

Chen, P.-Y., Wu, S.-y., & Yoon, J. (2004). The impacts of online recommendations and consumer ratings on sales. In *Proceedings of the 25th annual international conference on information systems.* (p. 711-723). WA, USA: AISeL.

Cheung, C. M. K., & Thadani, D. R. (2010). The effectiveness of electronic word-of-mouth communication: A literature analysis. In (p. 329-345). Bled, Slovenia: AISeL.

Chevalier, J. A., & Mayzlin, D. (2003). *The effect of word of mouth on sales: Online book reviews* (Tech. Rep.). National Bureau of Economic Research.

Clemons, E. K., Gao, G. G., & Hitt, L. M. (2006). When online reviews meet hyperdifferentiation: A study of the craft beer industry. *Journal of Management Information Systems*, *23*(2), 149-171.

Dewan, S., & Ramaprasad, J. (2007). Impact of blogging on music sales: The long tail effect. In *Workshop on information systems and economics.*

Duan, W., Gu, B., & Whinston, A. B. (2008, November). Do online reviews matter? an empirical investigation of panel data. *Decision Support Systems*, *45*(4), 1007–1016.

Etzion, H., & Awad, N. (2007). Pump up the volume? examining the relationship between number of online reviews and sales: Is more necessarily better? Melbourne, Australia: AISeL.

Hao, Y. Y., Ye, Q., Li, Y. J., & Cheng, Z. (2010, January). How does the valence of online consumer reviews matter in consumer decision making? differences between search goods and experience goods. In *System sciences (HICSS), 2010 43rd hawaii international conference on* (pp. 1–10).

Huang, A. H., & Yen, D. C. (2013). Predicting the Helpfulness of Online ReviewsA Replication. *International Journal of Human-Computer Interaction*, *29*(2), 129–138.

Hwang, Y. (2014). Understanding the different influences of online trust on loyalty by risk takers and avoiders. *International Journal of Human-Computer Interaction*, *30*(12), 977–984.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica: Journal of the Econometric Society*, 263–291.

Khare, A., Labrecque, L. I., & Asare, A. K. (2011, March). The assimilative and contrastive effects of word-of-mouth volume: An experimental examination of online consumer ratings. *Journal of Retailing*, *87*(1), 111–126.

Lee, S.-H. (2009). How do online reviews affect purchasing intention? *African Journal of Business Management*, *3*(10), 576-581.

Lelis, S., & Howes, A. (2011). Informing decisions: how people use online rating information to make choices. In *Proceedings of the 2011 annual conference on human factors in computing systems* (p. 2285-2294). NY, USA: ACM.

Liu, B. (2010). Sentiment analysis: a multifaceted problem. *IEEE Intelligent Systems*, *25*(3), 76–80.

Liu, Y. (2006). Word-of-mouth for movies: Its dynamics and impact on box office revenue. *SSRN eLibrary*.

Metzger, M. J. (2007). Making sense of credibility on the web: Models for evaluating online information and recommendations for future research. *Journal of the American Society for Information Science and Technology*, *58*(13), 2078-2091.

Nelson, P. (1970, March). Information and Consumer Behavior. *Journal of Political Economy*, *78*(2), 311–329.

Ozakca, M., & Lim, Y.-K. (2006). A study of reviews and ratings on the internet. In *CHI '06 extended abstracts on human factors in computing systems* (p. 1181-1186). NY, USA: ACM.

Paolacci, G., Chandler, J., & Ipeirotis, P. (2010). Running experiments on amazon mechanical turk. *Judgment and Decision Making*, *5*(5).

Park, D.-H., & Kim, S. (2008). The effects of consumer knowledge on message processing of electronic word-of-mouth via online consumer reviews. *Electronic Commerce Research and Applications*, *7*(4), 399–410.

Park, D.-H., & Lee, J. (2008). eWOM overload and its effect on consumer behavioral intention depending on consumer involvement. *Electronic Commerce Research and Applications*, *7*(4), 386–398.

Park, D.-H., Lee, J., & Han, I. (2006). Information overload and its consequences in the context of online consumer reviews. In *PACIS 2006 proceedings* (p. 28). Taiwan: AISeL.

Sen, S., & Lerman, D. (2007). Why are you telling me this? an examination into negative consumer reviews on the web. *Journal of Interactive Marketing*, *21*(4), 76-94.

Shaw, A. D., Horton, J. J., & Chen, D. L. (2011). Designing incentives for inexpert human raters. In *Proceedings of the ACM 2011 conference on computer supported cooperative work* (pp. 275–284).

Bum Chul Kwon is a postdoctoral researcher in the department of Computer Science and Information Science at University of Konstanz, Germany. His main research area is human factors and computational supports in information visualization and visual analytics. Kwon received his Ph.D. in Industrial Engineering from Purdue University in 2013.

Sung-Hee Kim is a postdoctoral researcher in the computer science department at University of British Columbia. Her main focus of her research is applying information visualization techniques to improve individuals' decision-making processes. She received her Ph.D. in the School of Industrial Engineering at Purdue University in 2014.

Timothy Duket received both his Bachelor and Master of Science in Industrial Engineering from Purdue University. His research interests range across many aspects of industrial engineering, most recently focusing on production systems and optimization. He is currently working as an engineering consultant at Bastian Solutions.

Adrián Catalán is an educator, tinkerer, software developer, Innovation Lab at GalileoU lead, and Android Google Developer Expert. He has worked as researcher in HCI (Human-Computer Interaction) and CCI (Crowd-Computer Interaction) areas. He has taught undergraduate and graduate master courses at several universities in Guatemala.

Ji Soo Yi is an associate professor in the School of Industrial Engineering at Purdue University. His research interests are on applying information visualization to decision making and developing relevant theories. Yi received his Ph.D. in Industrial and Systems Engineering from Georgia Institute of Technology in 2008.

List of Figures

Figure 1: F-distribution indicates the shape of distribution of reviews across five star ratings.

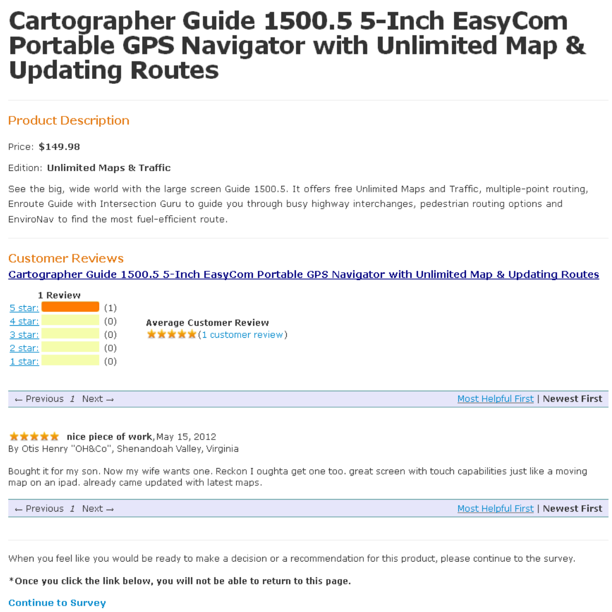


Figure 2: A screen shot of the experiment website with a review.

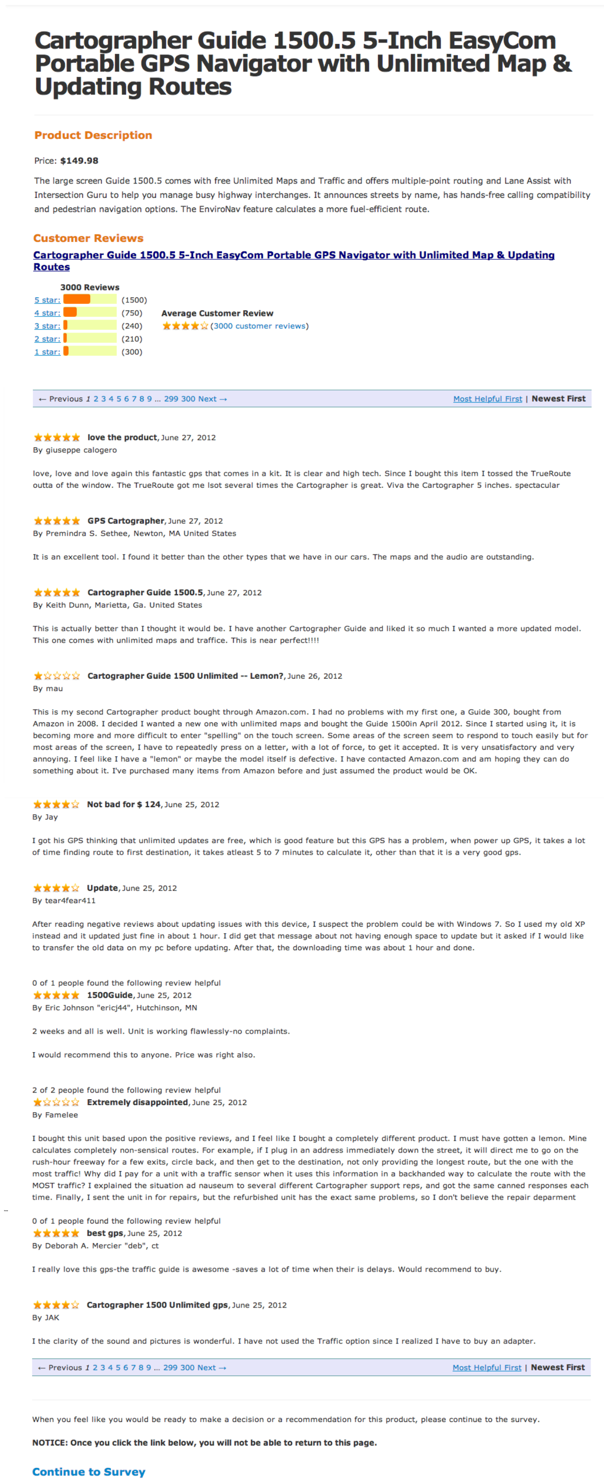
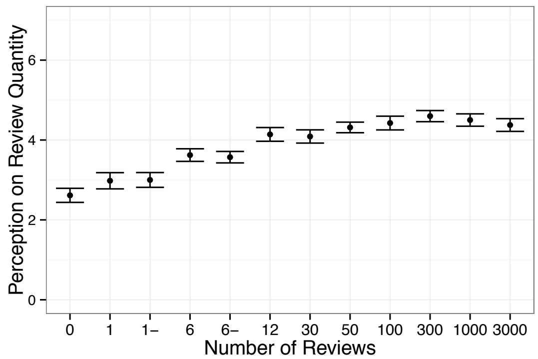
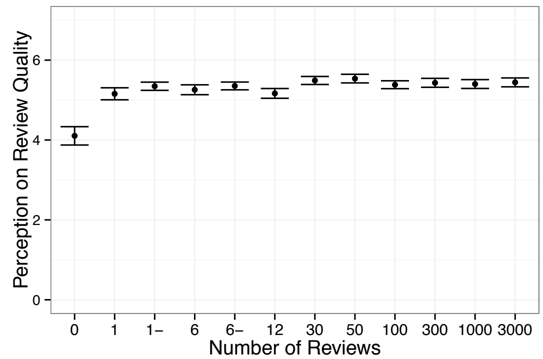


Figure 3: A screen shot of the experiment website with 3000 reviews.



(a) The interval plot of perception on review quantity



(b) The interval plot of perception on review quality

Figure 4: Interval plots of perceived review quantity and quality in Study 1.

Figure 5: Contribution levels of six information sources for Study 1.

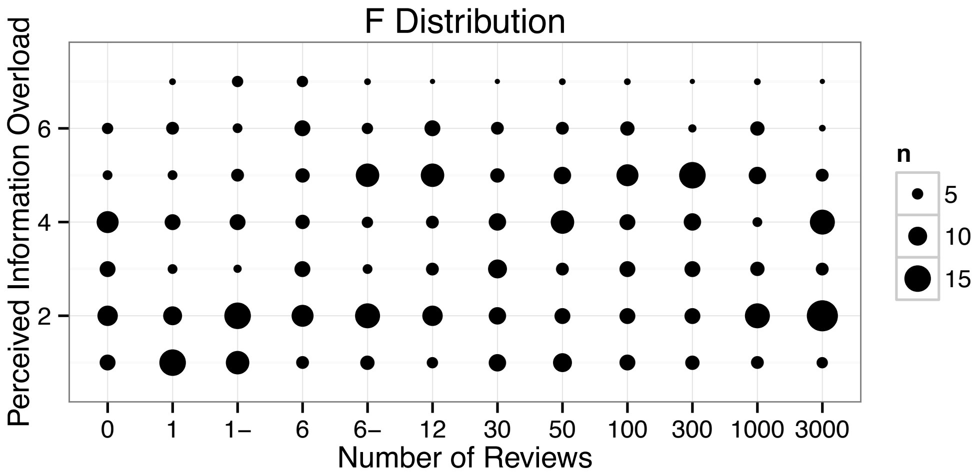


Figure 6: A bubble chart of perceived information overload versus number of reviews. The size of each bubble indicates the number of people who repsponded the corresponding rating (i.e., perceived information overload).

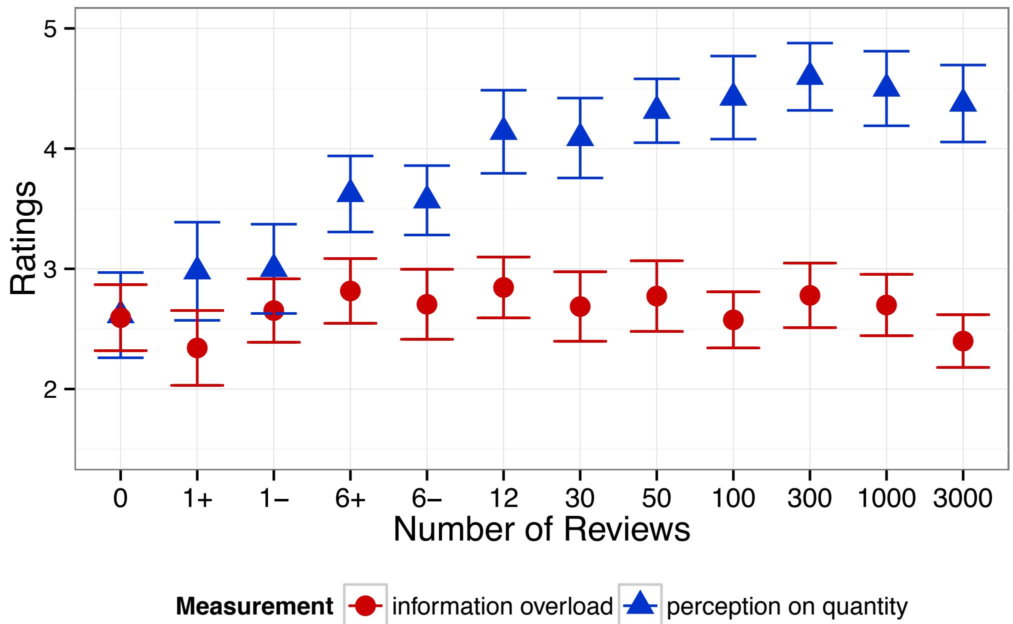


Figure 7: An interval plot of comparison of perceived information overload and perception on review quality.

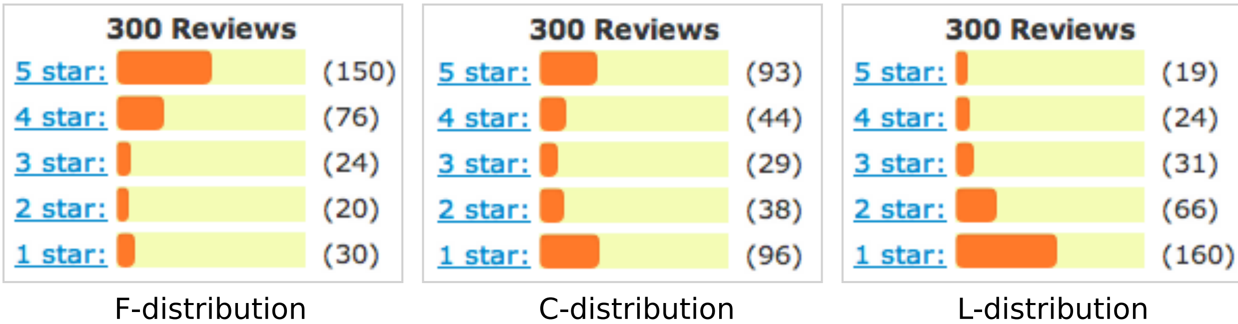


Figure 8: A screen shot of the three distributions of ratings shown when the number of reviews is 300.

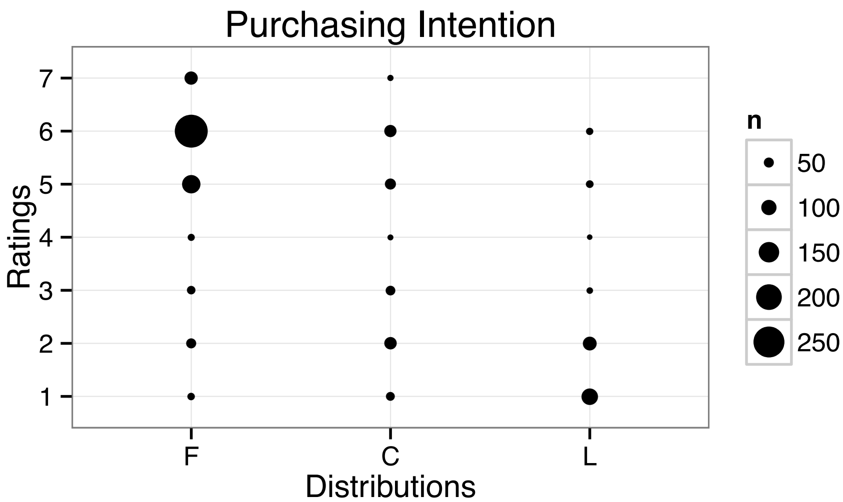


Figure 9: A bubble chart for purchasing intention for each distribution. The size of the circle indicates the number of participants.

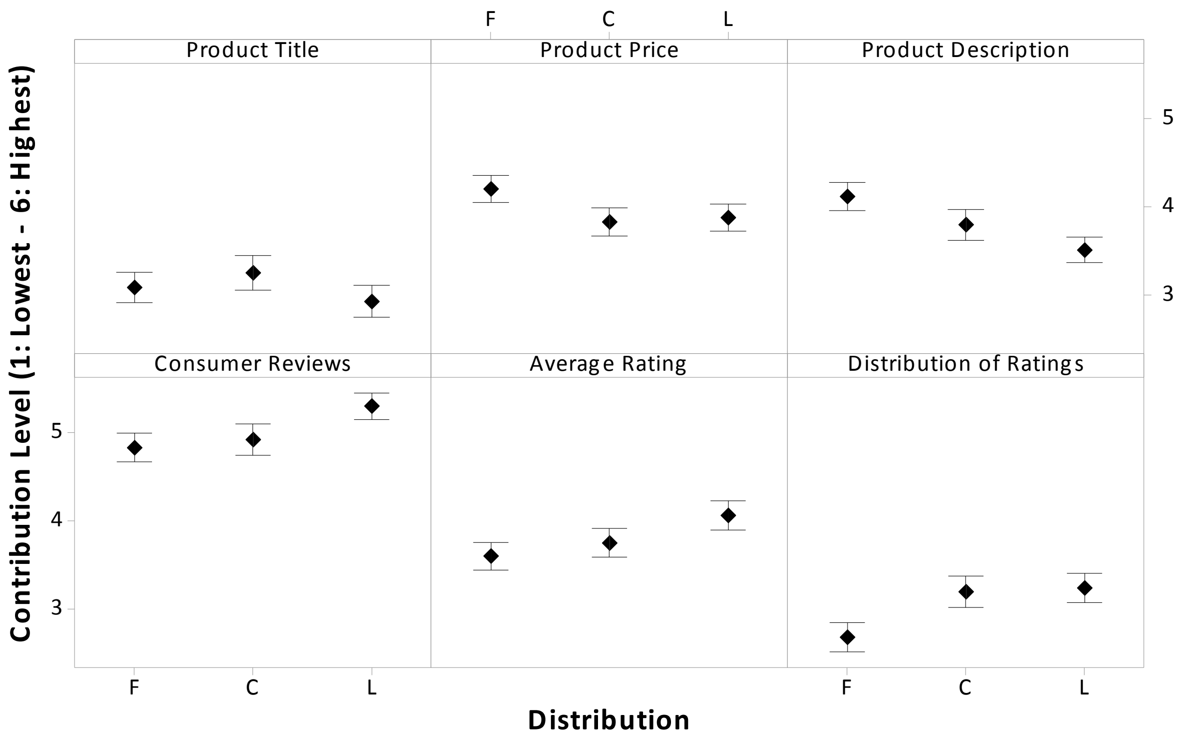


Figure 10: Contribution levels of six information sources for Study 2.

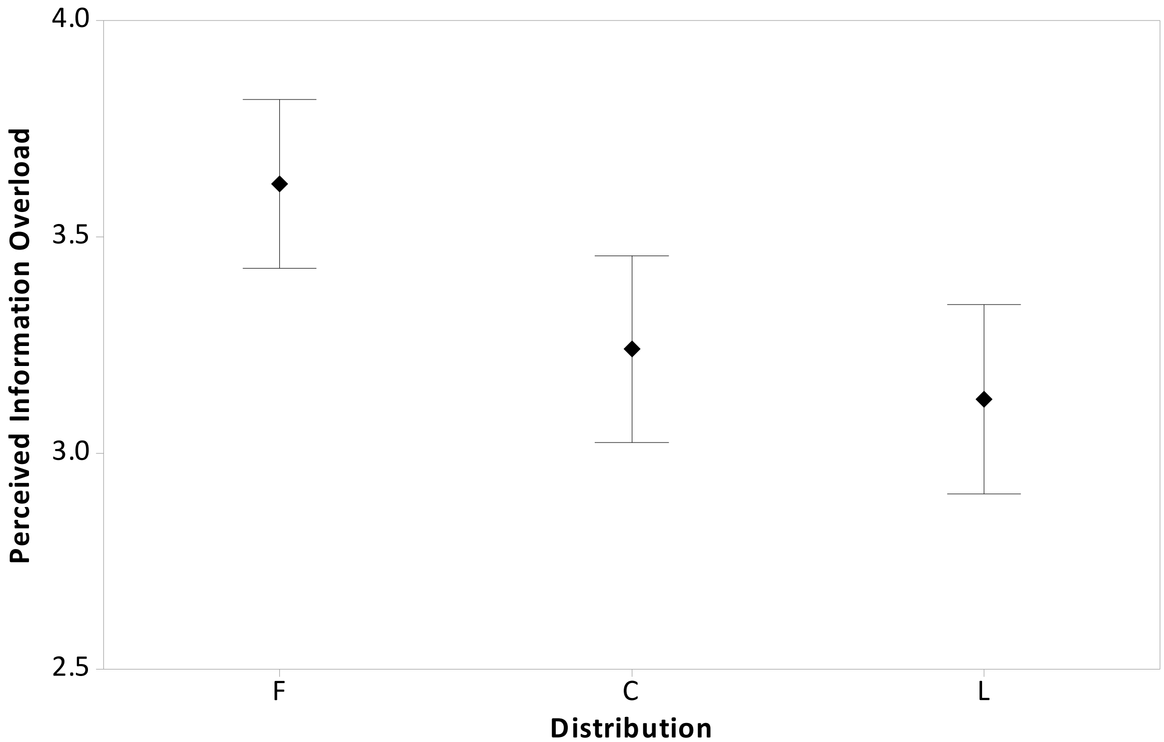


Figure 11: Information overload versus distributions.

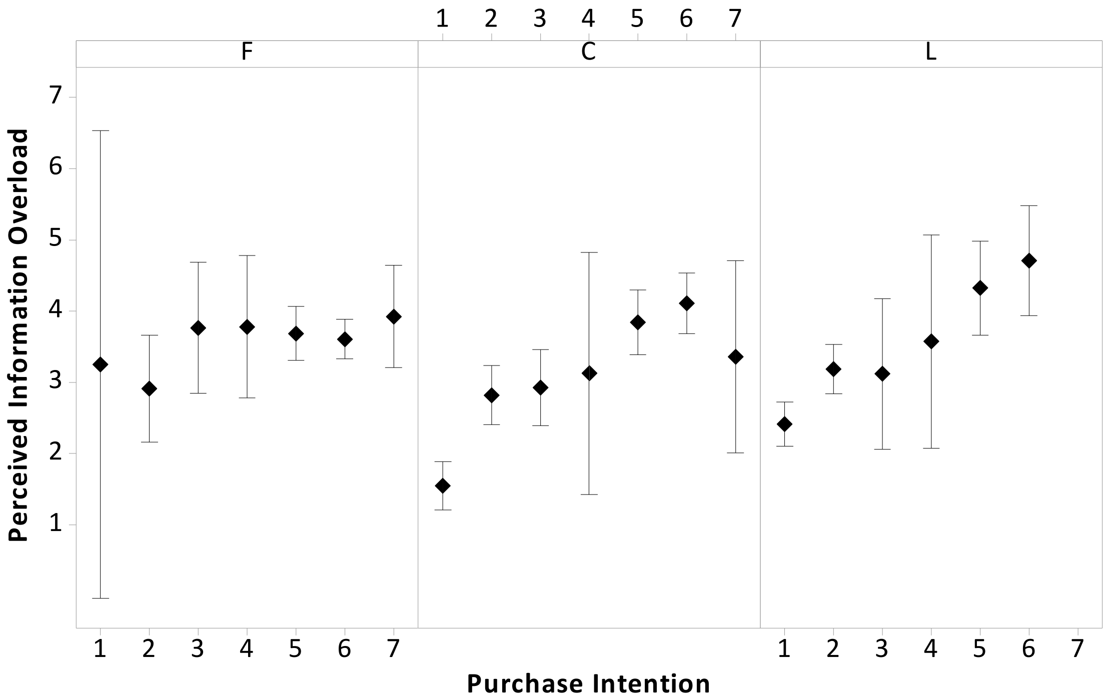


Figure 12: Perceived information overload versus purchasing intention.

List of Tables

Table 1: The Twelve Levels of Reviews in Study 1.

The Twelve Levels	Original	0	1+	1-	6+	6-	12	30	50	100	300	1,000	3,000
5-star	1,545 (50%)	0	1	0	3	3	6	15	25	50	150	500	1,500
4-star	789 (26%)	0	0	0	2	2	3	8	13	25	75	250	750
3-star	247 (8%)	0	0	0	1	0	1	2	4	8	24	80	240
2-star	207 (7%)	0	0	0	0	0	1	2	3	7	21	70	210
1-star	300 (10%)	0	0	1	0	1	1	3	5	10	30	100	300

Table 2: Statements in Survey Questionnaire.

Measures	**Original Survey Statement**	**Survey Statement Modified for Study 1**
Purchasing Intention (Park et al., 2006)	"How likely is it that you will buy this product?" "How likely is that you will recommend this product to your friends?"	I would recommend this product to Vina.
Information Overload (Park et al., 2006)	"How satisfied are you with your decision?"	I am satisfied with my decision.
	"How confident are you in your choice?"	I am confident that I made a wise decision.
	"How confused did you feel while performing this task?"	I felt confused while performing the task.
	"Do you feel that you suffer from information overload?"	I experienced information overload during the task.
Perceived Quality (Lee, 2009)	"Are these online reviews credible?"	The reviews that I read were credible.
	"Do these online reviews present sound arguments?"	The reviews that I read presented sound arguments.
	"Do these online reviews provide facts in support of their position?"	The reviews that I read provided facts to support their position.
Perceived Quantity (Lee, 2009)	-	I felt that there were too many reviews available.
	-	I would have preferred to have more reviews to read.

Table 3: Kruskal-Wallis Test Results on the Review Number versus Contribution Rankings of Information Sources.

Six Information Sources	Kruskal-Wallis Results
Product Title	$\chi^2(11) = 36.85$; $p < 0.01$
Product Price	$\chi^2(11) = 20.78$; $p = 0.03$
Product Description	$\chi^2(11) = 28.65$; $p < 0.01$
Consumer Reviews	$\chi^2(11) = 48.96$; $p < 0.01$
Average Rating	$\chi^2(11) = 20.57$; $p = 0.04$
Distribution of Rating	$\chi^2(11) = 9.77$; $p = 0.55$

Table 4: Number of Interactions Performed across the Twelve Levels of Review Volume in Study 1.

Interaction	0	1+	1-	6+	6-	12	30	50	100	300	1,000	3,000
Page Navigation	0	0	0	0	0	22	34	16	8	4	10	20
Sort	0	5	5	2	1	7	8	4	8	1	5	6
Filter by Star-rating	0	1	2	0	1	1	1	3	5	1	11	6

Table 5: Kruskal-Wallis Test Results on the Distributions of Ratings versus Contribution Ranking of Information Sources.

Six Information Sources	Kruskal-Wallis Results
Product Title	$\chi^2(2) = 4.85$; $p = 0.08$
Product Price	$\chi^2(2) = 14.77$; $p < 0.01$
Product Description	$\chi^2(2) = 28.82$; $p < 0.01$
Consumer Reviews	$\chi^2(2) = 23.46$; $p < 0.01$
Average Rating	$\chi^2(2) = 18.35$; $p < 0.01$
Distribution of Rating	$\chi^2(2) = 29.38$; $p < 0.01$

Table 6: Number of Interactions for C-Distribution Across the Twelve Levels of Review Volume in Study 2.

Interaction	6	12	30	50	100	300
Page Navigation	0	12	13	11	5	9
Sort by Date or Helpfulness	3	3	1	4	5	6
Filter by Star-rating	1	0	1	4	1	8

Table 7: Number of Interactions for L-Distribution Across the Twelve Levels of Review Volume in Study 2.

Interaction	6	12	30	50	100	300
Page Navigation	0	16	21	15	14	21
Sort by Date or Helpfulness	3	3	0	2	11	4
Filter by Star-rating	0	0	0	1	7	4