Evaluating content quality and helpfulness of online product reviews: The interplay of review helpfulness vs. review content

Article *in* Electronic Commerce Research and Applications · May 2012

DOI: 10.1016/j.elerap.2011.10.003

CITATIONS
309

READS
6,570

3 authors, including:

Nikos Korfiatis
University of East Anglia
77 PUBLICATIONS **1,265** CITATIONS

SEE PROFILE

Evaluating content quality and helpfulness of online product reviews: The interplay of review helpfulness vs. review content[1].

Nikolaos Korfiatis
(korfiatis@em.uni-frankfurt.de)
Institute of Informatics, Goethe University Frankfurt,
Frankfurt, Germany

Elena García-Bariocanal, Salvador Sánchez – Alonso
({elena.garciab,salvador.sanchez}@uah.es)
Department of Computer Science, University of Alcala
Alcala de Henares, Madrid, Spain

Cite this paper as:

Korfiatis, N., Barriocanal-Garcia, E, Sanchez, S. (2012). Evaluating content quality and helpfulness of online product reviews: The interplay of review helpfulness vs. review content. *Electronic Commerce Research and Applications*, 11(3), 205-217.

---

[1] The authors are grateful for helpful comments and suggestions received by Miguel-Angel Sicilia, Daniel Rodriguez and the participants in the seminars at the University of Alcala de Henares, Madrid, Spain (2008), the Copenhagedn Business School (CBS), Copenhagen, Denmark (2009) and the Economics Department of the University of Copenhagen. In this paper, we extend a similar study with a smaller dataset published in: Korfiatis N., Rodriguez D. & Sicilia M.A *"The impact of Readability on the Usefulness of Online Product Reviews: A Case Study on an Online Bookstore"*. WSKS 2008, pp. 423-432. Lecture Notes in Computer Science Vol.5288 Springer, Berlin/Heidelberg. The material in this earlier paper has been revised and extended to address, among other concerns and comments from the reviewers as well as the intuition behind the paper's theoretical basis, reasoning and results.

Running Title: Evaluating content quality and helpfulness of online product reviews: The interplay of review helpfulness vs. review content.

Abstract

Online reviews have received much attention recently in the literature, as their visibility has been proven to play an important role during the purchase process. Furthermore, recent theoretical insight argue that the votes casted on how helpful an online review is (review helpfulness) are of particular importance, since they constitute a focal point for examining consumer decision making during the purchase process. In this paper, we explore the interplay between online review helpfulness, rating score and the qualitative characteristics of the review text as measured by readability tests. We construct a theoretical model based on three elements: conformity, understandability and expressiveness and we investigate the directional relationship between the qualitative characteristics of the review text, review helpfulness and the impact of review helpfulness on the review score. Furthermore, we examine whether this relation holds for extreme and moderate review scores. To validate this model we applied four basic readability measures to a dataset containing 37,221 reviews collected from Amazon UK, in order to determine the relationship between the percentage of helpful votes awarded to a review and the review text's stylistic elements. We also investigated the interrelationships between extremely helpful and unhelpful reviews, as well as absolutely positive and negative reviews using intergroup comparisons. We found that review readability had a greater effect on the helpfulness ratio of a review than its length; in addition, extremely helpful reviews received a higher score than those considered less helpful. The present study contributes to the ever growing literature on on-line reviews by showing that readability tests demonstrate a directional relationship with average length reviews and their helpfulness and that this relationship holds both for moderate and extreme review scores.

Keywords: Product reviews, Readability Tests, Review Helpfulness, Word of Mouth

1. Introduction

Use of the Web as a source of information has undisputedly affected several areas of human activity as regards how market transactions take place, being no exception (Li & Bernoff, 2008). The move towards a read-write Web and the ability to provide more relevant information to consumers on online marketplaces has been also demonstrated to influence the product choice process in electronic markets (Dellarocas, 2003, Godes et al., 2005). In particular, the most profound contribution in this respect has been the ability to present and promote customers' opinions of the products that they have purchased (Clemons, Barnett, & Appadurai, 2007, Clemons, Gao, & Hitt, 2006). As a result, a vast quantity of online reviews are available to consumers, which they can use to become better informed about the product or service that they are considering purchasing. Furthermore, it is argued in the literature that these reviews are sometimes more trustworthy than traditional sources in the printed press (Z. Jiang & Benbasat, 2004, 2007, Senecal & Nantel, 2004). This has led to attempts to extract useful information automatically from the review text (Zahn, Loh and Liu, 2009).

Online reviews in general constitute third-party evaluations by consumers of the product or services advertised on a website, and are displayed next to the product description in order to enhance customer perception and improve the perceived communication characteristics of the medium (Kumar & Benbasat, 2001). Their availability has been proven to have a positive effect on the volume of online sales (Archak, Ghose, & Ipeirotis, 2011; Chen, Xie, & Hall, 2008; Chevalier & Mayzlin, 2006). Websites that present an increasingly large number of user-supplied reviews tend to experience an increased volume of sales, especially in the case of products that are not of a tangible nature, as happens with more traditional bricks-and-mortar sales channels (Wolfinbarger & Gilly, 2001). On-line reviews exert even more influence in the case of products whose utility can only be evaluated upon consumption (for example, books). Intuitively, it is clear that this is the case for experience goods.

Nelson (Nelson, 1970) defined the concept of an experience good as a product or a service where the quality and utility for a consumer can only be determined upon consumption. This implies that in order for consumers to decide to consume or purchase this product or service, they must rely on previous experiences which provide an indicator of whether this product or service is worthy of purchase or not. One example of an experience good is a book: the utility to the consumer of reading a particular book can only be judged after reading that book. Therefore, with experience goods such as books, the producers (e.g. the publishers) often use reviews written by authoritative sources, such as literature experts, to provide an opinion and endorse the book, so that those consumers who trust

these sources will purchase this product.

The issue of trust becomes more apparent in online review literature when it is related to the effect of the review source (Pavlou & Dimoka, 2006). For example, in the case of books, publishers often choose to include a short review by leading reviewers (e.g. the NY Times bestsellers column) on the back cover, in order to attribute the review to a trusted source. However, it can be argued that this ploy is affected by reporting bias, since publishers are unlikely to include "unexciting" reviews on the back-cover (Zhang, Ye, Law, & Li, 2010). Similarly, reporting bias can also occur on online reviews. Hu et al. (Hu, Pavlou, & Zhang, 2006) have shown that online reviews related to books or CDs present a U-shaped distribution, with the extremes anchored to either "very good" (five stars) or "very bad" (one star) ratings. This creates the need to apply some kind of filter to those reviews that are justified by the textual description provided by the consumer. This is mainly displayed as "review helpfulness" on online marketplaces. For example, online retailers such as Amazon frame review helpfulness as a bimodal choice, asking the users: "*Was this review helpful for you?"*

From a cognitive perspective, review helpfulness can be used as a conformity filter. For example, reviews closer to consensus may be considered more helpful by potential consumers than those exhibiting extremes of opinion (J. Jiang, Gretzel, & Law, 2010). Another effect to consider is that reviews by individual consumers often express a personal view of their experience with the product, and this may differ from the expectations of the interested buyer. For example, whereas one person may have expected the book to contain more action elements, another might not be interested in that specific characteristic. In this case, the textual characteristics of the review act as a filter between the expectations of one consumer and the actual review posted by another. This raises several questions concerning the quality of the reviews themselves, namely, the relationship between review helpfulness and product utility for the reviewer who posted the review on the online marketplace. For example, do issues of conformity arise when someone reports extreme opinions ("very good" or "very bad") but provides a clear justification for their opinion? These issues indicate the need for further analysis of the review text and how it relates to the usefulness expressed by users.

Consequently, in this study we explored the qualitative or textual characteristics of online consumer reviews. A preliminary assumption was that submitted reviews reflected consumers' experience of product use. Therefore, the underlying assumption was that the individual submitting the review had consumed the product and was in a position to report his or her own personal experience and/or judgment of the product, since he/she had prior experience of the product (negative or positive, depending on the value of the review rating). Secondly, we considered the review to act as a

"justification" for the rating, enabling the potential buyer to evaluate whether the review was fair or not. This was reflected in the text submitted together with the rating provided. In addition, online marketplaces provide the possibility of review meta-rating, enabling interested buyers to evaluate how helpful a particular review was during the product selection process. Once again, in this case the review text acts as the main source for other consumers when evaluating the helpfulness of a specific review.

In this study, we focused our attention not on the numerical rating of the review, but rather on the review text itself. In particular, we sought to evaluate how the style and comprehensibility of a review, as appraised by computerized assessment, might affect the helpfulness of a review – in other words, the number of people who found it helpful out of the total number of people who had read and evaluated the review. In order to investigate this issue, we employed readability metrics applied to a dataset of reviews together with their helpfulness evaluations, collected from the bookstore section of Amazon U.K[2]. An earlier study by Korfiatis et al. (Korfiatis, Rodríguez, & Sicilia, 2008) discussed the relationship between review readability, or the style used to write the review, and the helpfulness score it received. This earlier study has been taken as the point of departure for developing a more extensive theoretical discussion of the relationship between readability score and review helpfulness.

In order to address this issue, we based our study on some preliminary assumptions. First, we focused on the viewpoint of consumers who were interested in buying a product or service from an online marketplace, and their consideration of whether these codified pieces of information were helpful or not. Secondly, we hypothesized that in addition to considering the review score given by a particular review, consumers would also evaluate the importance of the review according to how similar it was to their own communication codes, denoted by the way in which the review was written. This can be assessed by the qualitative characteristics of the review text as measured by readability tests.

Our findings indicate that review helpfulness is affected by writing style, and in particular, that the stylistic elements employed seem to have a much greater influence than the extensiveness of the text on its resulting helpfulness score. Furthermore, we find that even for extreme helpfulness score values, both the stylistic elements and extent of the review text had a great influence on the

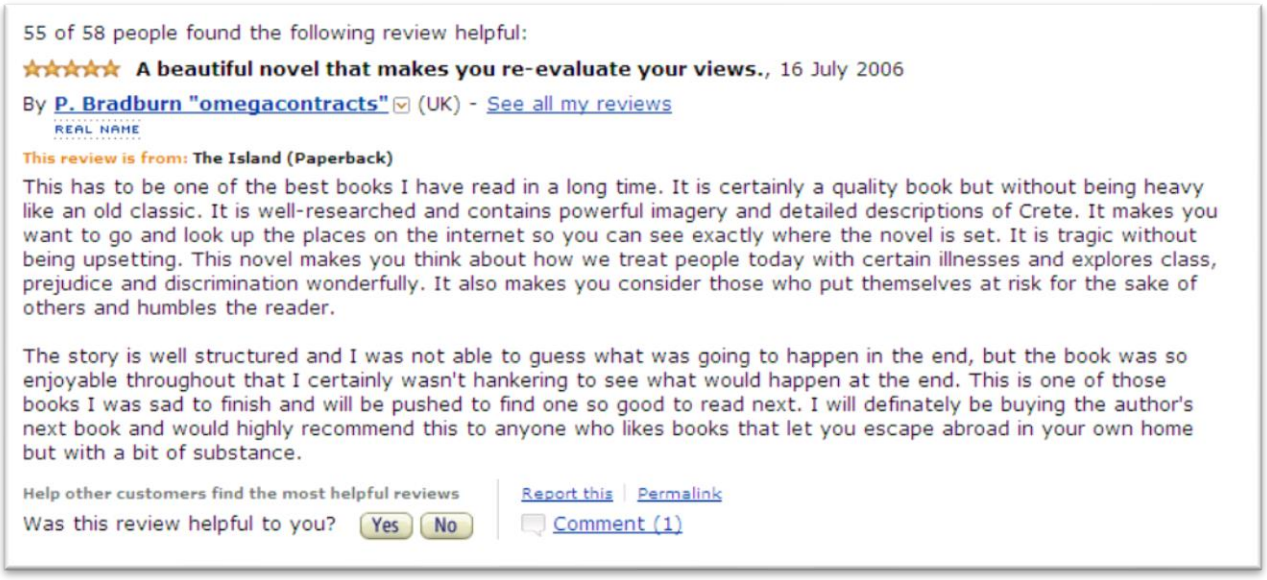
[2] http://www.amazon.co.uk/

helpfulness ratio that the review received. An inter group comparison showed that this also holds for extreme helpfulness ratio values and rating scores.

To this end, this paper is structured as follows: Section 2 provides the theoretical reasoning and motivation behind this study. Section 3 provides a background on readability tests and how they are calculated, together with the meaning of their norms. Section 4 presents the operationalization of the constructs discussed on section 2 and an analysis of the dataset we collected. We discuss the results and their alignment with existing literature in Section 5. Finally, Section 6 summarizes the conclusions of this study and its limitations, and indicates directions for future research.

2. Theoretical Model and Reasoning

One important feature of the online review is the way in which it is displayed to the consumer. Online marketplaces such as Amazon have invested a considerable amount of resources in designing and improving the store interface, since online store interaction plays a significant role in the selection process (Eroglu, Machleit, & Davis, 2003, Van der Heijden & Verhagen, 2004).


55 of 58 people found the following review helpful:

★★★★★ **A beautiful novel that makes you re-evaluate your views.**, 16 July 2006

By P. Bradburn "omegacontracts" (UK) - See all my reviews
REAL NAME

This review is from: **The Island (Paperback)**

This has to be one of the best books I have read in a long time. It is certainly a quality book but without being heavy like an old classic. It is well-researched and contains powerful imagery and detailed descriptions of Crete. It makes you want to go and look up the places on the internet so you can see exactly where the novel is set. It is tragic without being upsetting. This novel makes you think about how we treat people today with certain illnesses and explores class, prejudice and discrimination wonderfully. It also makes you consider those who put themselves at risk for the sake of others and humbles the reader.

The story is well structured and I was not able to guess what was going to happen in the end, but the book was so enjoyable throughout that I certainly wasn't hankering to see what would happen at the end. This is one of those books I was sad to finish and will be pushed to find one so good to read next. I will definately be buying the author's next book and would highly recommend this to anyone who likes books that let you escape abroad in your own home but with a bit of substance.

Help other customers find the most helpful reviews | Report this | Permalink

Was this review helpful to you? Yes No | Comment (1)

Figure 1: A typical review interface obtained from Amazon UK

Usually, a review interface contains three main elements: (a) The number of stars given by the reviewer (commonly known as *review rating*), (b) The review content, which is displayed in the body of the text along with review meta characteristics such as the author's pseudonym (user name or nickname) and the date of submission, and (c) The helpfulness of the review displayed in a textual way such as: *X out of Y people found the following review helpful*. Figure 1 provides an example of the review display interface from Amazon UK used in this study.

As mentioned above, review helpfulness represents the number of helpful votes that the review has received out of the total number of votes that have been given regarding the helpfulness of the review. One particular issue concerning the way the helpfulness score is constructed is the fact that most of the helpful or unhelpful votes come from consumers who are in the middle of their purchase process; therefore, the helpfulness score can be interpreted as an aid in the purchase decision. Consequently, it can be theorized that the helpfulness score reduces uncertainty about product quality, an important issue on online marketplaces (Zhu & Zhang, 2010). For example, Pavlou et al. (Pavlou, Liang, & Xue, 2007) theorized the importance of uncertainty, examining the question from the perspectives of both the buyer and the seller on an online marketplace. Following on from this approach to uncertainty, we based our theoretical model on three specific aspects which arise throughout the review evaluation process: *Conformity*, *Understandability* and *Extensiveness*. Our theoretical model is depicted in Figure 2. These aspects and their relationships to the hypotheses presented in the model are discussed in the sections that follow.

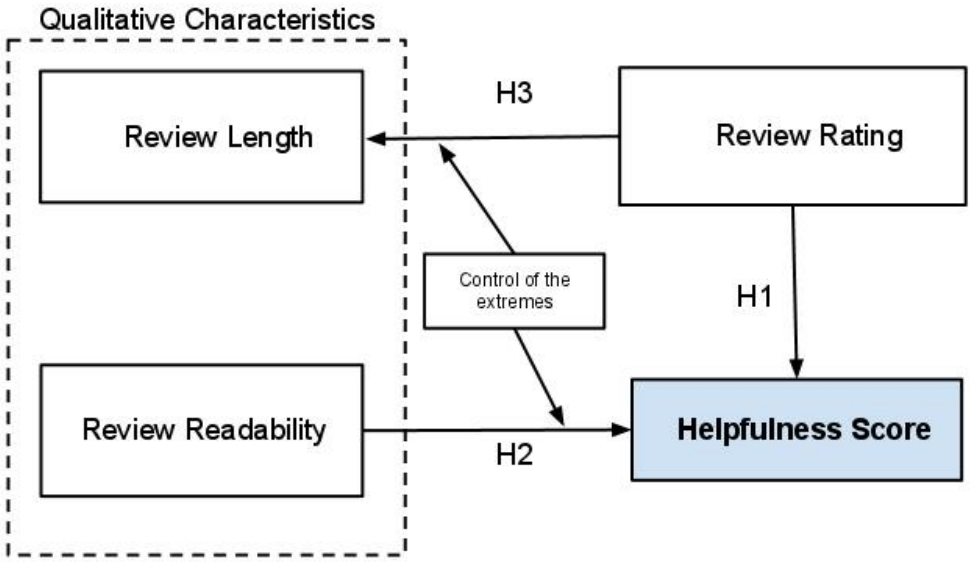


Figure 2: The model used in this study.

2.1 Conformity

In the context of online reviews, conformity refers to the idea that a review will be considered more helpful when the rating it gives is close to the consensus reached among ratings on this product. From a social psychology point of view, this idea has its roots in the theory of conformity (Kelman, 1958, Moscovici, 1985), whereby an individual's point of view and beliefs are conditioned by group consensus regarding perceptions. A very positive or a very negative review might not be considered helpful inasmuch as it does not reflect the consensus reached by other reviews available

on this product. This can be expressed in two ways: either by not voting at all, or voting the review as unhelpful. This implies the possibility of reporting bias; however, we worked on the assumption that reporting bias would be minimal since the risk of the transaction could also be considered minimal (Jurca, Garcin, Talwar, & Faltings, 2010).

Bearing the foregoing in mind, we hypothesize the following:

Hypothesis 1a (Conformity): *The helpfulness of a review will be directly affected by its rating*

Since the review process uses a five-point star scale, it enables many reviewers who wish to express indifference to give a three-star rating (middle option) to a product. For example, in their analysis of online reviews for books on Amazon, Forman et al. (Forman, Ghose, & Wiesenfeld, 2008) found that indifferent ratings (around three stars) were considered less helpful compared to extreme ratings (one star / five stars). In other words, one-sided reviews (referring only to either positive or negative aspects) were considered much more helpful than moderate reviews, which reported both positive and negative aspects. Furthermore, Pavlou and Dimoka (Pavlou & Dimoka, 2006) also found that extremely positive or negative ratings of eBay sellers were assessed as more informative than moderate ratings. This latter finding may be explained by the interplay between purchase intention, which has a binary outcome (to purchase or not), and the formation of a consideration set (Shocker, Ben-Akiva, Boccara, & Nedungadi, 1991). A product review that presents a one-sided argument (in favor of or against purchase) is considered more helpful in a search process, since it eliminates or strengthens the position of the product with regards to the list of alternatives or items in a consideration set. Parra and Ruiz (Parra & Ruiz, 2009) also found evidence that online settings provide instruments enabling consumers to view alternatives quickly and thus reduce the size of the consideration set in an online shopping scenario. Thus, we made the assumption that online reviews also act as a second-stage instrument for screening alternatives in the consideration set formed by the consumer, and that in this case, extreme reviews containing one-sided messages would be favored over moderate reviews containing two-sided messages.

Although this may appear to be a logical development of the theoretical viewpoint, other approaches in the literature have suggested the opposite. For example, an early study by Crowley and Hoyer (Crowley & Hoyer, 1994) asserted that two-sided messages (in this case reviews reporting positive and negative aspects) were more persuasive than one-sided positive reviews which reported finding no negative aspects. In addition, Eisend (Eisend, 2006) suggested that two-sided messages can also enhance source credibility, since the consumers' perception of these will be that of an unbiased third party evaluator (as in the case of consumer advocacy groups). Therefore, generalizing from this

theoretical viewpoint, we may also posit that:

Hypothesis 1b (Conformity): *The helpfulness of a review is unaffected by its rating*

Although conformity is one aspect which explains why a review is considered helpful, other aspects, namely review understandability and expressiveness, should also be explored. These two constructs concern the actual textual content of the review itself and are discussed below.

2.2 Understandability and Expressiveness

Understandability defines the level of comprehension that a piece of text or a report (in our case, a product review) requires in order to be understood and/or to make an informed decision when using it as input (Smith & Taffler, 1992). Specifically, the understandability of a review text is directly related to its qualitative characteristics, such as readability and length. In other words, readability is operationalized on how easy it is to read and comprehend a piece of text containing judgments related to the product being evaluated. Related to H1, a more understandable piece of review text containing subjective evaluations (where product attribute-related information is excluded) will be considered more helpful than a more difficult to understand one. This can be theorized at the level of cognitive effort and, in particular, in terms of the review text's cognitive fit to an average consumer with a normal level of expertise regarding the product evaluated (Park & Kim, 2009). Theoretically, when the information expressed in the text matches the consumer's own information-processing strategy, a cognitive fit occurs (Vessey & Galletta, 1991). Furthermore, understandability plays a major role in how the numerical evaluation of the review is justified, and is thus helpful in assisting the consumer to screen a product for a purchase in conjunction with the theoretical intuition discussed in Hypothesis 1. Therefore, we hypothesized that understandability of a review text would directly affect the helpfulness of the review (the more understandable the text, the more helpful the review) and theorized that this would also be the case for both extremely helpful and extremely unhelpful reviews.

Hypothesis 2a (H2a): *The helpfulness of a review is directly affected by its qualitative characteristics, and in particular, by the readability of the review text.*

Hypothesis 2b (H2b): *The extreme helpfulness of a review is directly affected by its qualitative characteristics, and in particular, by the readability of the review text.*

While helpfulness is one aspect where the review text may serve as a proxy, the review rating is

another which plays an ex-ante role in determining what the review text might look like. That is, while qualitative aspects may affect review helpfulness, the effect of the star rating given by the review on the formation of the text's qualitative characteristics must also be taken into account. In particular, the extent to which the review text is expressive when justifying a rating should be considered.

One possible indicator of expressiveness could be the star rating (Tsang & Prendergast, 2009). When a customer rates a product, he/she is required to rate his/her experience of the product using a single dimension indicator, the star rating. If the rating given alongside the review tends towards an extreme (thus communicating a one-sided message), this procedure is easier since for the the review text to be consistent, it only needs to contain the positive or negative experience. On the other hand, if the review is moderate (communicating a two-sided message), then it should contain a mixture of arguments, making it longer than an extreme-rated review. This is also related to review consistency, in that where the review text offers an explanation for the rating, the merit of a highly subjective rating can be evaluated by the consumer herself. (Gerdes, Stringam, & Brookshire, 2008, Stringam & Gerdes, 2010).

Therefore, we theorize that the length of the review text would be affected by the rating given and that this would also hold true for extremely positive and extremely negative reviews:

Hypothesis 3a (H3a): *The length of the review text is affected by the value of the rating given by the review.*

Hypothesis 3b (H3b): *The length of the review text is affected by the extreme value of the rating given by the review.*

While Hypothesis 2 relates the qualitative characteristics (readability) of the review text to review helpfulness, Hypothesis 3 postulates that the review rating is a generator of the review text which directly affects readability. We can hypothesize that expressive/lengthier reviews will have different qualitative characteristics to shorter reviews, since the level of cognitive effort will be in accordance with the length of the review text.

Having described the theoretical model and intuition behind this study, we will now describe the readability tests that we used to operationalize our constructs and to evaluate our model.

3. Background on readability tests

In general terms, the concept of readability describes the effort and the educational level required for

a person to understand and comprehend a piece of text (DuBay, 2004, Zakaluk & Samuels, 1988). In more formal terms, a readability test is a formula produced by applying linear regression to subjects, regarding the reading ease of different pieces of text that these had been asked to comprehend using specific instruments. The purpose of a readability test is to provide a scale-based indication of how difficult a piece of text is for readers to comprehend, according to the linguistic characteristics of that text. Thus, a readability test can only provide an indication of how understandable a piece of text is based on its syntactical elements and style. Most of the readability tests in the literature indicate the school grade level required in order to comprehend the piece of text provided. In addition to current applications in the field of education, readability tests have also been applied to situations where the subject is required to read a piece of information in order to make a decision, or to comprehend the logic behind the present outcome, for example, in the case of accounting reports (Smith & Taffler, 1992).

Readability Measure	Score Range	Measurement Implications
Gunning's Fog Index[DP1]	1-12[3]	Indicates the educational grade level required. The lower the grade, the more readable the text
Flesch Reading Ease index	0-100	Scores above 40% indicate that the text is understandable by literally everyone. As the value of the index decreases, the comprehensibility of the text becomes more difficult.
Automated Readability Index	1-12[1]	Indicates the educational grade level required. The lower the grade, the more readable the text
The Coleman-Liau Index	1-12	Indicates the educational grade level required. The lower the grade, the more readable the text

Table 1: The readability tests used in this study

It is fair to say that the attention a review might receive from interested buyers is a large extent associated with its readability. In this case, a readability test assessment of a review can provide an indication of whether someone who had evaluated a particular review as helpful was actually able to comprehend the piece of text that was submitted. On the other hand, it may also be the case that where some reviews were not considered helpful, this may have been influenced by the readability of

[3] In order to better understand the variability between the three readability tests, in the present study we normalized both the ARI and the Fog index to the 100 metric of the Flesch index.

the content.

However the use of a readability test presents some major drawbacks which it was necessary to take into consideration when analyzing the results of this study. In particular, the result of a readability formula cannot tell us whether the review content expresses personal opinions of the product and/or contains gender, class or even cultural bias. To avoid differences in cultural background and ensure language proficiency, we only collected reviews from the U.K. store of an online marketplace, in order to restrict our study as far as possible to native English speakers located in one geographical region (population from one country)[4].

Readability tests have been used to study the qualitative characteristics of several types of text in different areas of information science, and a large set of readability indexes has been developed over the years (Paasche-Orlow, Taylor, & Brancati, 2003). For our study, we selected four major readability tests which have been used extensively to evaluate the readability of a piece of text by individuals with various educational levels.

One particular reason for considering readability tests an ideal instrument for evaluating the helpfulness of an online review stems from the theoretical basis presented in the introduction and the previous section. Specifically, we made the assumption that any evaluation of review helpfulness was to a large extent dependent on the justification for the review star rating given in the review text. In previous research in the area of education, it has been reported that text that is easily readable will be understood better; therefore those who understand a piece of text better will form a more justified opinion about it[5].

Table 1 lists the readability tests used in the present study. These include the Gunning's Fog index (FOG), the Flesch/Kincaid Reading Ease Index (FK), the Automated Readability Index (ARI) and the Coleman-Liau Index (CLI). All four tests evaluate the readability of a text by consistently breaking the text down into its basic structural elements, which are then combined using an empirical regression formula. However, it is important to note that not all the indexes measure the same characteristics. The Fog and Coleman-Liau indexes measure complexity, whereas the Flech/Kincaid

[4] Furthermore, the readability tests used in the analysis section of this paper have only been developed for the English language. However, proficiency in English undoubtedly affects their validity. This should be taken into account since Amazon.UK also has customers from other European countries.

[5] UK national readability survey. See http://www.hmie.gov.uk/documents/publication/hmiear03-05.html

and ARI indexes measure reading ease. This is important in terms of analysis, since complexity and comprehension might be different depending on the context in which the text is written.

Another important limitation concerning readability tests is that they can only be used to evaluate short texts (which makes them ideal for evaluating texts such as online reviews), since a reader's comprehension of a text also involves cognitive properties which are beyond the scope of this study (e.g. sentiment analysis). The reasoning behind the calculation and the norms of these instruments is described in the sections below. It should be noted that each readability test uses a set of constants and coefficients for calculating the readability score. These constants have been computed by the creators of the test and do not depend on the data. In particular, the numbers in each formula are the result of a regression-like evaluation of a subset of the standard population, and apply solely to the English language (based on language entropy and character frequency in each word).

3.2 The Gunning-Fog Index (FOG)

Gunning's Fog index (Gunning, 1969) represents a measure of the extent to which an individual with an average high school education would be able to comprehend the evaluated piece of text. The following equation describes the empirical relationship in the Fog Index:

$$FOG = 0.4 \times \left(\frac{Words}{Sentence} + 100 \times \left(\frac{N(complex\_words)}{N(words)} \right) \right)$$

An obvious difficulty in measuring the Fog index for a given text is the evaluation of the number of complex words. In our analysis, we considered a word as complex wherever it had more than two syllables.

3.3 The Flesch–Kincaid Reading Ease (FK)

The Flesch–Kincaid Reading Ease index (Flesch, 1951, Kincaid, Fishburne Jr, Rogers, & Chissom, 1975) uses a core linguistic measure based on syllables per word and words per sentence in a given text. The Flesch test is used to evaluate the complexity of the text in order determine the number of years of education which would be needed for someone to understand the text being assessed. The following equation describes how the Flesch-Kincaid score is calculated for a given text:

$$FK = 0.39 \times \left(\frac{total\_words}{total\_sentences} \right) + 11.8 \times \left(\frac{total\_syllables}{total\_words} \right) - 15.59$$

The variables *total_words*, *total_sentences* and `total_syllables` denote the total number of

words, sentences and syllables found in the text, respectively. For calculating the Flesch score of a particular review we broke the text down into sentences, then into words and finally into syllables which were combined using the constants presented in the formula above. It can easily be implied from the mathematical expression given that the fewer the number of words per sentence, the better the Flesch test readability score.

3.4 The Automated Readability Index (ARI)

The Automated Readability Index (ARI) differs from the Gunning-Fog and the Flesch-Kincaid in the use of more simple metrics to evaluate the readability of a typical English language text. In order to calculate the ARI for a given review, we first calculated the total number of characters (excluding standard syntax such as hyphens and semicolons) and the total number of words.

$$ARI = 4.71 \times \left(\frac{characters}{words} \right) + 0.5 \times \left(\frac{words}{sentence} \right) - 21.43$$

Calculating the ARI involved the same steps as for the Fog and Flesch indexes, but in addition it was also necessary to calculate the number of characters, or review length. The ARI can provide an indication of the influence of review length on review readability.

3.5 The Coleman-Liau Index (CLI)

The Coleman-Liau Index (Coleman & Liau, 1975) is similar to the Automated Readability Index, except that the second part of the formula considers a more careful selection of the textual characteristics of the piece of text assessed. The CL index has been developed specifically for machine-based scoring, thus the calculations that it involves are quite lengthy when conducted manually. The following formula describes the Coleman-Liau index:

$$CLI = 5.89 \times \left(\frac{characters}{words} \right) - 0.3 \times \left(\frac{sentences}{word} \right) - 15.8$$

Index calculation considers fragments of sentences of 100 words multiplied by a constant (0.3).

4. Data, Methods and Results

Having described our theoretical model and provided a background to the readability tests used to operationalize the measurement of stylistic elements in the review text, we will now present an analysis of the reviews in our dataset.

4.1 Data Collection and Operationalization of Variables

In order to collect the data and test our theoretical model, we developed a web crawler to capture the contents of the ***books section*** on Amazon UK. The crawler consisted of two parts: (a) A web client, to randomly pick items from the front page of the bookstore and (b) a client of the web service interface provided by Amazon Web Services (AWS), where the data for each item was collected[6]. The list of books was stored on a relational database which we used in order to obtain the reviews available on each individual product page. The data was collected in April-March, 2008. Books which had been published more than six months before the time of data collection, or which had no rating, were omitted from the database. Furthermore, we excluded books on special offer or offered at a discount, in order to control for price effects (Pavlou & Dimoka, 2006).

Variable Code	**Variable description**
`productid`	The ID of the product that the review referred to. This was used to control for the publication date and other product characteristics.
`summary`	The summary / title of the review.
`content`	The actual content of the review. Used for content analysis.
`rating`	The rating that the review gave, measured on a 1-5 Likert scale.
`totalvotes`	The number of total votes that were given to the review.
`helpfulvotes`	The number of "helpful" votes awarded to the review.

Table 2: The main variables of the initial dataset collected using the web crawler

Amazon UK was selected in order to obtain a dataset which avoided language comprehension heterogeneity among reviewers and consumers, which might have had a negative influence on text comprehension. This was important since readability tests are pointless when a reader is not a native speaker of the language in which the text is written, since many languages differ in syntactical form and the writing style used in the review might be totally different from the reader's native language (thus the level of comprehension by non-native speakers might be affected).

[6] Amazon web services provides a REST api which returns an xml file containing the reviews of each product by supplying a unique identifier called ASIN. More information can be found in Amazon Web Services API documentation – http://aws.amazon.com.

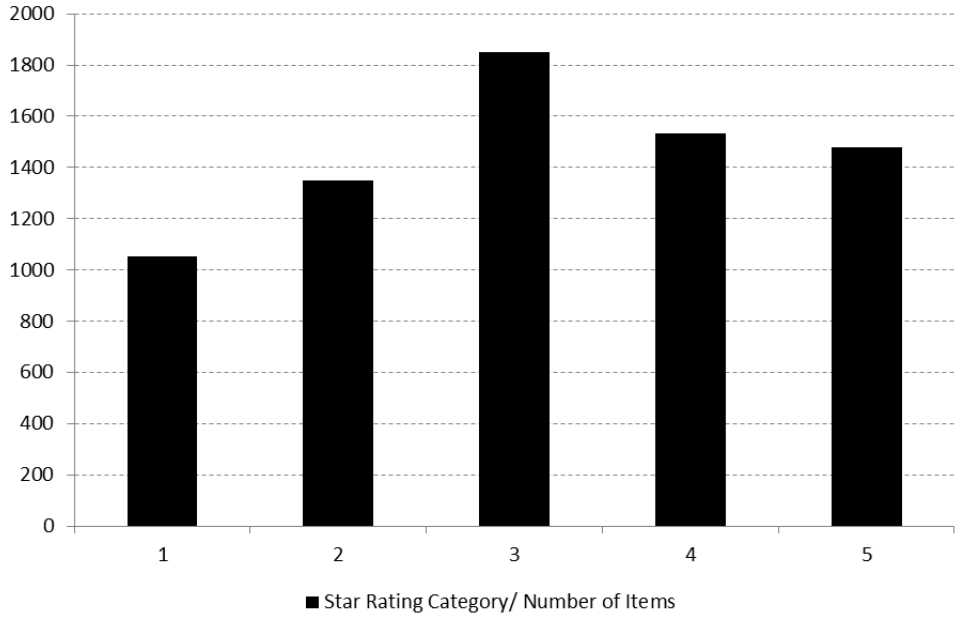
Table 2 provides a description of the variables used in our dataset. These variables can be categorized into two groups; numerical expressions of the review (`rating`, `totalvotes`, `helpfulvotes`), and the textual or qualitative characteristics (`summary`, `content`) of the review content.

One particular issue with the dataset collected was the need to by-pass promoted items such as bestsellers. Since these items were more accessible to visitors to the online bookstore/future customers, there was always the possibility of selection bias towards the more visible items, potentially resulting in higher exposure to more recent product reviews. To avoid this, the web crawler maintained a list of the frequency with which items were displayed on the front page, and randomly chose items listed by categories.

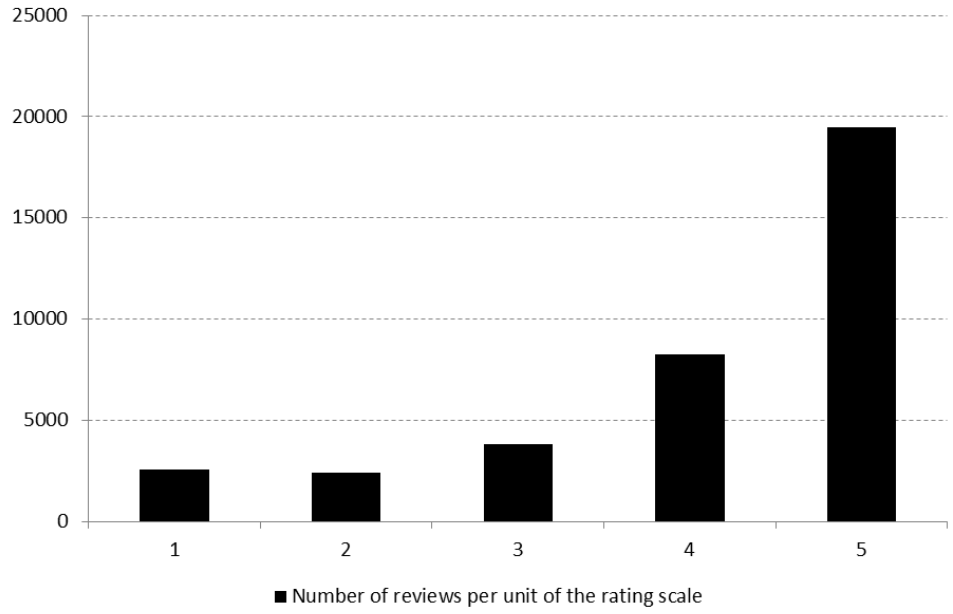
Figure 3 presents the distribution of the number of items per average star rating category (top) and the distribution of the number of reviews per star rating category (bottom). We define the helpfulness ratio of a review (HR) as the number of helpful votes a review received *(*`helpfulvotes`*)* divided by the total number of votes evaluating the helpfulness of that review *(*`totalvotes`*)*. Thus, the dependent variable for our analysis is defined as:

$$HR = \frac{helpfulvotes}{totalvotes}$$

The helpfulness ratio is a measure of the quality of the review according to the readers themselves. The greater the number of helpful votes a review received from those who had evaluated the review, the higher the helpfulness ratio.



(a)



(b)

Figure 3: *Distribution of the (a) average rating values among the items in our dataset and (b) the number of reviews per unit of the rating scale. (Total number of reviewed items/books: N_{books}=7262)*

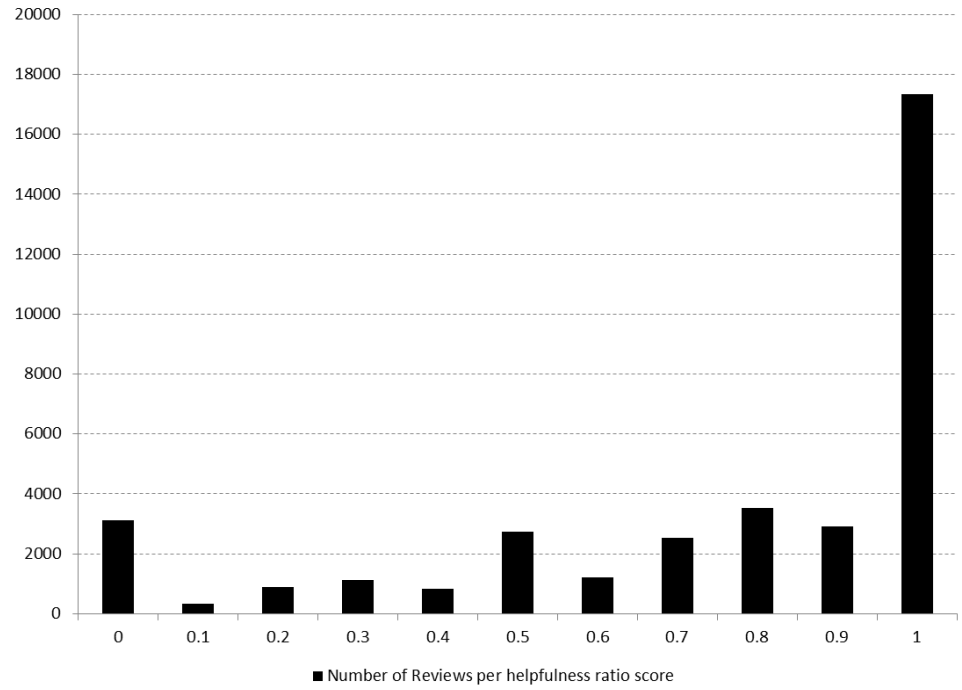


Figure 4: *The distribution of reviews on our dataset per decimal point of helpfulness score (N=36856)*

The dataset consisted of 36,856 reviews with a total votes value greater than zero (totalvotes>0), ensuring that the reviews on our dataset had been evaluated for their helpfulness at least once. Figure 4 provides an overview of the distribution of helpfulness scores on our dataset. It is interesting to note that around 47% (17,335) of the reviews had received a full score from the readers, indicating that around half of the reviews were very highly acclaimed by their readers. In other words, for this particular of group of reviews, the number of helpful votes was the same as the number of potential buyers who had read the reviews. On the other hand, we found that approximately 9% (3,119) of the reviews were judged to be totally unhelpful by their readers, receiving an absolute 0 of helpful votes.

One issue that we wish to address here is that length and readability can be coupled by the same concept of qualitative characteristics of the text, when both are computed by an algorithm (review length in number of words or characters, readability from a readability test) and not by using a panel of experts.

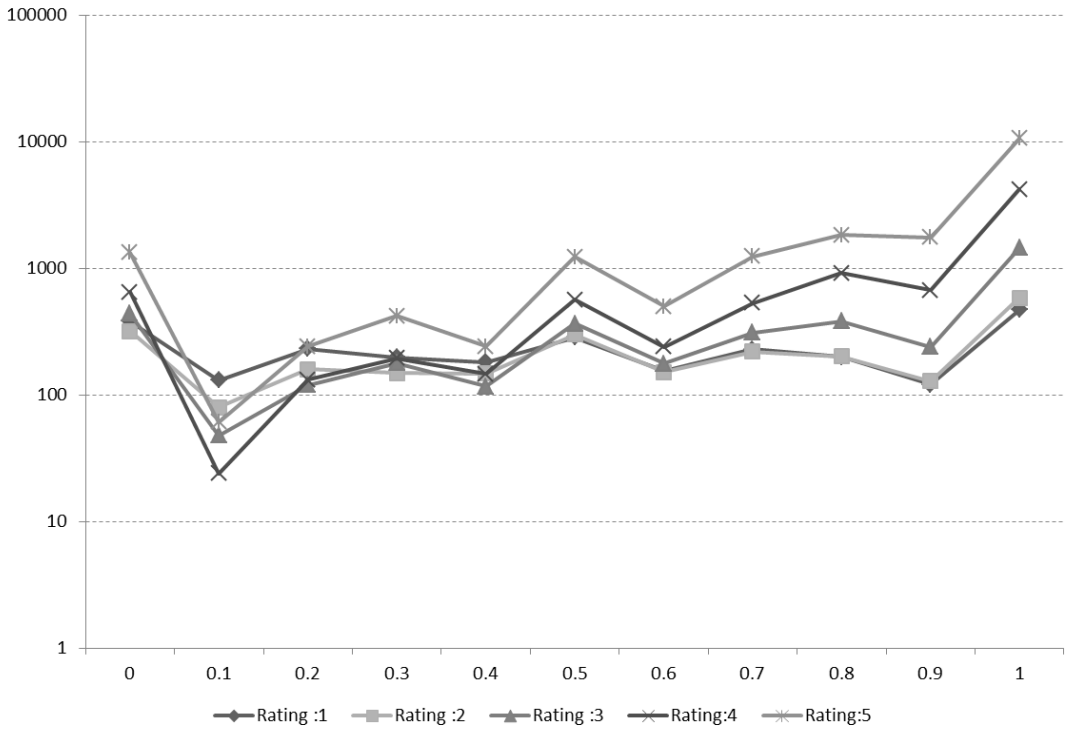


Figure 5: Distribution of the rating scores and average helpfulness in our dataset. Axis x depicts the value of the HR per 10% change. Axis y depicts the number of reviews and is on logarithmic scale. As expected, the distribution follows a U-shape function. Absolutely positive reviews are considered more helpful than moderate reviews.

Figure 5 presents the distribution of the rating scores that were given to all reviews on our dataset per decimal point of the helpfulness ratio. It is interesting to note that more than 70% of the reviews were highly positive (rating>=3), indicating a tendency for positive extremes to be voted more helpful by consumers. This is confirmed by the theory described in the introduction. The variable we used in order to apply the readability tests was the *review's content*, represented on the website by the content and summary variables, respectively. To obtain the results of the readability formulas described in Section 2, we used the "style" command option of the GNU-Dict package[7].

[7] The version of the GNU Dict that is currently available is 0.7 and can be downloaded from http://www.gnu.org/software/diction/diction.html

Variable	**Mean**	**Std. Dev.**	**Min**	Max
Helpfulness Ratio	.750	.322	0	1
Helpful votes	5.03	7.63	0	230
Total Votes	6.61	9.16	1	372
Rating	4.08	1.23	1	5
Number of Words	160.07	135.84	74	1854
CLI	8.55	2.44	0	28.5
ARI	26.41	3.40	0	263
FOG	.99	.010	0	214
FK	9.81	6.25	0	206

Table 3: *Descriptive statistics for the variables used to estimate the helpfulness of a review* (N=37221)

As mentioned earlier, the helpfulness of a review was measured as the number of helpful votes cast divided by the total number of votes regarding the helpfulness of that review. Table 3 provides the descriptive statistics for the main variables on our dataset. The average review rating was 4 (see also Figure 5), with an average text length of 160 words. The helpfulness ratio had a mean value of 0.750 or 75.0%, indicating that three out of four reviews were considered helpful by visitors to the website. On average, a review received almost two (9.16-7.63 = 1.53) negative votes out of the average nine votes evaluating how helpful the review was. As described previously, we considered the total number of votes to represent the minimum number of consumers who had read this review and / or considered it in their selection process for the particular item in question.

4.2 Analysis and Results

Table 4 presents the inter-correlation matrix that we obtained from running Spearman non parametric inter-item correlations between the items in our dataset. The correlation coefficients were obtained by conducting a pair- wise correlation between the variables at a 1% ($P<0.01$) level of significance. It is interesting to note that the four readability tests used showed a high inter-item correlation, which can be explained by the fact that they were used to evaluate the same piece of text.

		1	2	3	4	5	6	7	8	9
1	hr	1								
2	Total Votes	-.150**	1							
3	Helpful Votes	.246**	.879**	1						
4	Number of Words	.148**	.246**	.313**	1					
5	Rating Score	.244**	-.059**	.072**	-.062**	1				
6	FK	.087**	.149**	.185**	.437**	-.070**	1			
7	CLI	.100**	.121**	.163**	.309**	-.031**	.608**	1		
8	ARI	.089**	.147**	.184**	.433**	-.068**	.985**	.648**	1	
9	FOG	.082**	.141**	.175**	.410**	-.068**	.979**	.598**	.962**	1

Table 4: Spearman non-parametric inter-correlation matrix between the constructs in our dataset. (**P < 0.01).

Before conducting a causal evaluation of the constructs that affected the HR of a review, we ran a non-parametric correlation on our sample in order to evaluate the significance of the interrelationships between the different constructs used in the analysis. **Table 4** provides the results of the non-parametric Spearman inter-correlation between the items in our dataset. Since the HR presented an underlying robust and significant relationship with the other constructs ($p<0.01$), we proceeded to model the effects of the other constructs on the HR.

As described in Section 2 on the theoretical basis, we expected that the HR would be affected by readability-related constructs. Therefore, we wanted to investigate (a) the effect of considering stylistic elements on the predictive power of the HR model, and (b) whether this effect would remain unchanged regardless of the number of people who had read the review. That is, we wished to explore the effect of readability when considering both extreme reviews and high numbers of users who had evaluated a particular review. Another issue to consider, related to Hypothesis 1, was whether reviews closer to the extremes would be considered more helpful than moderate ones, implying a concave relationship.

To this end, we theorized the following model variations:

Model 1: $HR = \beta_{11}*rating + \beta_{12}*rating^2 + \beta_{13}\,helpfulvotes + \beta_{14}*totalvotes + \beta_{15}*wordcount + \varepsilon_1$

We wanted to explore the effect on the model fit when the stylistic elements of the review text were added. Therefore we considered:

Model 2: $HR = \beta_{21}*rating + \beta_{22}*rating^2 + \beta_{23}\,helpfulvotes + \beta_{24}*totalvotes + \beta_{25}*wordcount + \beta_{26}*FK + \beta_{27}*CLI + \beta_{28}*AR + \beta_{28}*FOG + \varepsilon_2$

Given the concave rating relationship for the HR, we further examined whether a concave relationship existed between the totalvotes and helpfulvotes and the HR after inclusion of the readability metrics. Therefore, we obtained:

Model 3: $HR = \beta_{21}*rating + \beta_{22}*rating^2 + \beta_{23}\,helpfulvotes + \beta_{24}*totalvotes + \beta_{25}*wordcount + \beta_{26}*FK + \beta_{27}*CLI + \beta_{28}*AR + \beta_{28}*FOG + \beta_{29}*helpfulvotes^2 + \beta_{29}*totalvotes^2 + \varepsilon_2$

A critical issue at this stage of the analysis was selection of the regression model to be used since, by definition, the dependent variable has limited low and high extremes. With the aim of assessing the fit of the above models, we used a variance of regression analysis, specifically, TOBIT regression. The TOBIT model is particularly applicable to this analysis as it is widely used in econometrics for estimating the effects of independent variables on a non-negative dependent variable. Furthermore, Mudambi and Schuff (Mudambi & Schuff, 2010) have suggested in a previous study that TOBIT is

well-suited for this purpose since (a) people are inclined to vote on extreme reviews and (b) longer reviews will receive more votes.

Table 5 presents the regression output of the TOBIT regression for the three different models. We used Efron's R-square and the likelihood ratio as an assessment of fit, which are reported by default on a stata tobit procedure as pseudo r-square and LR respectively.

	Model 1		**Model 2**		**Model 3**	
	Coef.	Std. err.	Coef.	Std. err.	Coef.	Std. err.
(Constant)	.4996045***	0.011	.4165113***	0.014	.5192255***	0.013
Rating	.0546365***	0.007	.0476397***	0.007	.015095*	0.007
$Rating^2$	-0.0015232	0.001	-0.0003728	0.001	.0022829*	0.001
Helpfulvotes	.0451459***	0.001	.0447567***	0.005	.0746766***	0.001
Totalvotes	-.033151***	0.000	-.0329146***	0.004	-.0533867***	0.001
# of Words	.0002408***	0.000	.0001884***	0.001	.0001209***	0.000
FK			**.022807*** **	0.003	.0203454***	0.003
CLI			**.0109902*** **	0.000	.0087728***	0.001
AR			**-.013513*** **	0.002	-.0127597***	0.002
FOG			**-.0051302***	0.002	-0.0036836	0.002
$Helpfulvotes^2$					**-.0002923*** **	0.000
$Totalvotes^2$					**.0001435*** **	0.003
Efron's R^2	0.316		0.354		0.451	
Likelihood	*10664(df 5), p=.0000*		*10945(df: 9),p=.0000*		*15221(df:11),p=.0000*	

Table 5: Regression results of the helpfulness ratio (HR).
Significance: *p<0.05, **p<0.01, *** p<0.001, N_{Tobit}=36586.

It can be seen that the first model presented an acceptable level of predictive power (Model 1: R^2=.316), indicating that the model provided a good description of the dataset at this stage. One interesting aspect to note here is that no significance was found for the quadratic effect of the rating on the HR, perhaps due to the large number of positive reviews that existed on the dataset, rendering the concave relationship less significant. The number of words was significant but had no effect as a coefficient on the HR.

All three models supported Hypothesis 1, namely, that the review rating will have a significant effect on the HR. We also observed that this effect decreased slightly when considering the stylistic elements of the review text. When evaluating the second model, which contained the readability

metrics, we observed a notable increase in the fit of the model (Model 2: R^2=.354), and the readability score coefficients were higher than the word count coefficient. This finding, together with the values of the coefficients (β_{23},…, β_{28}), confirm H2, namely, that the qualitative characteristics of a review will have an effect on the HR. What is even more interesting is that all four coefficients presented a higher value than the word count, implying that readability had a greater effect than review length.

In other words, readability will influence the number of people who read the review, and therefore there will be heterogeneity among the different review text comprehension levels. By adding this quadratic relation to Model 3, we observed an increase in the model fit (Model 3: R^2=.451) of almost 20%. This implied that the quadratic effect of totalvotes and helpfulvotes, although not important in terms of coefficient size, nevertheless increased model fit. Having confirmed H2, we proceeded to examine the effect of extreme ratings (one-sided messages) by also assessing H3 using intra group comparisons.

4.3 Inter-group comparisons of the extremes and the impact of rating on review helpfulness

In order to test whether the characteristics of a review varied depending on the HR, we divided the dataset into different categories, based on the HR that the reviews received and the rating score that the reviews contained.

Specifically, the dataset was divided into a total of three group categories using the following criteria:

A. If the review's HR was lower or higher than 0.5 (the number of helpful votes the review had received was lower or higher than the number of readers who had read the review and did not consider it helpful);

B. If the review's HR was lower than 0.25 or higher than 0.75 (the number of helpful votes the review had received was less than one quarter or more than three quarters of the readers who had read the review and did not consider it helpful);

C. If the review's rating score was lower or higher than 3, which, considering the 5 value likert scale, divided the dataset into reviews that had received a low (1-2) or high (4-5) rating.

D. If the review's rating score was absolutely negative (1) and absolutely positive (5).

Since dividing the data into two groups provided two independent samples from the same dataset, we

used intergroup comparisons to determine the significance of the difference between the means of specific characteristics in these groups.

The different samples are tabulated in Table 6. Groups A and B were formed by taking the value of the HR as a split criterion, and groups C and D with the value of the rating score, respectively. Below, we will discuss whether the hypothesis held for a bimodal split, and then for the extreme quartiles of the HR (1st and 4th) and the lowest and highest rating score.

Group	Identifier	(#) of Observations	(%) of the dataset	Split Criterion
Groupid-A	A_1	5844	17.45	HR<0.5
	A_2	27641	82.55	HR>=0.5
Groupid-B	B_1	3980	10.88	HR<0.25
	B_2	22511	61.53	HR>0.75
Groupid-C	C_1	5025	13.73	Rating <3
	C_2	27730	75.79	Rating >=3
Groupid-D	D_1	2584	7.06	Rating=1
	D_2	19460	53.19	Rating=5

Table 6: Sampling and selection procedure for the groups used in our analysis.

4.3.1 The helpfulness of a review is affected by the extreme rating that the review has received

Having split our dataset into two grouping variables, we are able to test the relationship between the helpfulness ratio of a review (HR) and the extreme values of the rating that this review had received. In particular, we were interested in determining whether the helpfulness ratio presented any relationship to the rating that a review received, by comparing the means of the ratings.

As mentioned previously, we selected a non-parametric test to determine whether the mean rating value was the same across the groups presenting a high or low HR. For the grouping variable Groupid-A, the groups were divided equally by the HR (Groupid-A =1 if the HR was lower than 0.5, and Groupid-A =2 if the HR was higher than or equal to 0.5). We selected the Mann-Whitney

test to compare the mean rating value between the two groups. This particular test was selected because the Mann-Whitney test is a non-parametric statistical test that, unlike parametric tests (t-test), does not rely on the assumption of normality (the rating distribution among groups follows a normal distribution) and is unaffected by arbitrary sample sizes.

For the grouping category Groupid-A, we ran the Mann-Whitney test for a total of N= 33,485 observations. The Z value that we obtained from the test was Z =-39.407, with a p-value of p=0.000, which is highly significant at three degrees of freedom. Thus, we can reject the hypothesis that the rating mean will be the same for reviews with a high or low helpfulness ratio.

Group/Subgroup		*Number of Reviews*	*Average Rating*	*Std. Dev*
A	A1	5844	3.43	1.51
	A2	27641	4.25	1.08
B	B1	3980	3.47	1.41
	B2	22511	4.34	1.50

Table 7: Tabulation of average rating score for the two groups. N_{GroupA}=33305,N_{GroupB}=26491

In order to verify the above result, and determine whether the rating had no effect on the different HR values for the upper and lower limits (since most of the HR present values of 0 and 1), we used the second grouping (Groupid-B), which divided the dataset into two parts, namely, the first and fourth quartile of the HR values. We once again ran the Mann-Whitney test for a total of N= 26,884. The Z value obtained from the test was Z=-36.009, which corresponds to a p-value of P=0.000, indicating that the hypothesis that the rating will be the same between the two groups was false. Both results confirmed that the rating was affected by the HR of the review, even for extreme HR scores.

4.3.2 Extreme helpfulness of a review is affected by the qualitative characteristics of the review (readability).

In order to test whether the helpfulness of a particular review was affected by the qualitative characteristics of the review text, we employed the same procedure for both grouping categories (Groupid-A and Groupid-B), conducting a comparison of a representative readability test. Table 8 provides the tabulation of the groupings using the Flesch-Kincaid readability average as a comparison between the categories.

Group/Subgroup		*Number of Reviews*	*Average FK (1-12)*	*Std. Dev*
A	A1	5844	9.51	0. 91
	A2	27641	10.4	1.04
B	B1	3980	8.58	1.20
	B2	22511	10.04	1.94

Table 8: Tabulation of the Average FK score among the Groupings A and B. N_{GroupA}=33305,N_{GroupB}=26491.

Running the test for the first grouping category (Groupid-A), we obtained a Z value of Z= -12.580 and a p-value of *p=0.000,* which is again highly significant at three degrees of freedom. The hypothesis that average readability score value will be the same whether the HR is extremely high or extremely low is rejected, indicating that review text readability affects the review HR. Similar results were also obtained for the other readability tests.

4.3.3 The extensiveness of the review (review length) is affected by the extreme values of a review's rating score.

We examined whether the number of words was different when the rating was high or low. For this, we considered two groups and compared the average number of words. For this type of comparison (since the dependent variable was not categorical), we ran a two-tailed t test, not assuming equal variances. Table 9 provides the tabulation of the average number of words for both grouping categories.

Group/Subgroup		*Number of Reviews*	*Average Word Count*	*Std. Dev*
C	C1	5025	145	112.02
	C2	27730	161	139.64
D	D1	2584	133	103.41
	D2	19460	154	137.33

Table 9: Tabulation of the average number of words for the Groupings C and D. N_{GroupC}=32755,N_{GroupD}=22044

For the grouping category C, we obtained a t-value of t=- 7.7832, giving a p-value of *p=0.000*, which was highly significant at 32753 degrees of freedom. The same test for the grouping category Groupid-D provided a t value of t=-7.5182, again giving a p-value of p=0.000 for df=22042 degrees of freedom. Neither test assumed equal variances.

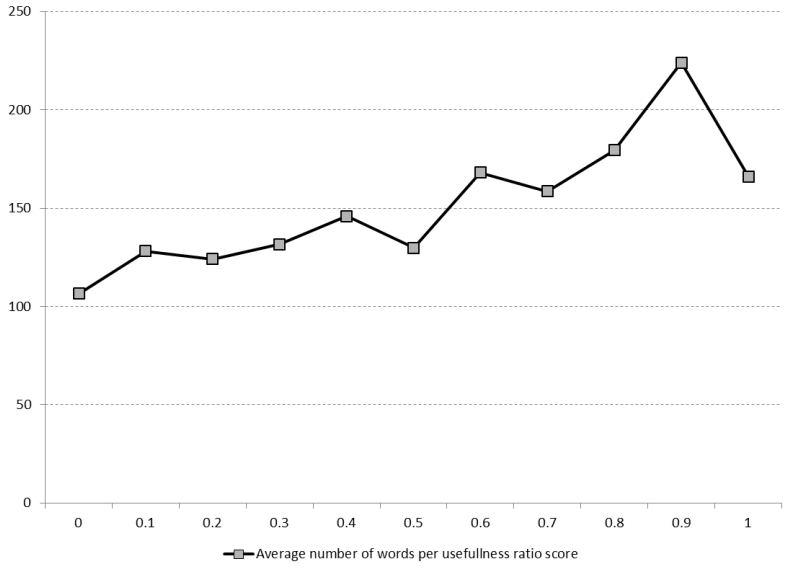


Figure 6: Distribution of the average length of the review text (number of words) according to the review rating

Figure 6 can be used to visualize review text trends across the rating scale values used when a review is submitted (1-5). It is clear that reviews with positive ratings tended to contain more text (as evidenced by the number of words contained in the review text).

5. Discussion

5.1 Validation of the theoretical model

The results of the tests, together with the interpretation of the regression coefficients that we obtained (see Section 4.2), have provided us with some interesting insights into the relationship between the helpfulness of an online review, according to consumers / visitors on an online marketplace, and the qualitative characteristics that the particular review presents. The findings in relation to the hypothesized theoretical model are summarized in Table 10.[DP2]

Specifically, from the data analysis reported above, we found support for the following:

a) High helpfulness of a review is affected by its positive rating value.

Looking at Table 7 (inferred from the results), a significant trend can be seen in the HR towards reviews with higher ratings, and this also holds true for extreme HR values. This may be explained by the fact that consumers (as visitors to an online information resource) tend to read appraisals of a product first (the fact that a review is marked with 5 stars also increases attention from a usability point of view). In accordance with the definition of the helpfulness ratio, this indicates that the higher

the review rating value, the higher the amount of helpful votes (future) customers will give to this particular review. In our dataset, reviews with a rating value above three had a higher number of helpfulness ratios with a perfect score, indicating that the higher the number of helpful votes the review received, the higher its helpfulness ratio. This latter implies that customers react to positive and negative reviews differently, confirming the findings of the study by Hu et al. (Nan Hu, L. Liu, & Jie Zhang, 2008).

Hypothesis	Description	Supported
H1a	The helpfulness of a review is directionally [DP3]affected by its rating.	*Yes*
H1b	The helpfulness of a review is non-directionally [DP4]affected by its rating.	*No*
H2a	The helpfulness of a review is directionally affected by its qualitative characteristics and in particular by review text readability.	*Yes*
H2b	The **extreme** helpfulness of a review is non-directionally affected by its qualitative characteristics and in particular by review text readability.	*Yes*
H3a	The length of the review text is affected by the review rating score value.	*Yes*
H3b	The length of the review text is affected by the **extreme** review rating score value.	*Yes*

Table 10: Summary of the findings.

b) Highly helpful and extremely helpful reviews contain more readable text than reviews that are less helpful or not helpful at all.

The word length and readability scores confirmed that the review justification (in this case, the text) provided an indicator of why a review was considered highly helpful by a consumer. That can be explained by the fact that consumers base their evaluation of a review on whether it is well justified and provides as much information as possible to the customer, enabling the latter to form his or her own opinion about the quality of the product in question (in our case, a book) and reduce uncertainty about its quality.

c) Reviews are longer when they are positive or absolutely positive and shorter when they are negative and absolutely negative.

Surprisingly enough, our results indicated that there was a clear relationship between the rating value a review gave and its qualitative characteristics in terms of review length. This was also confirmed for the lowest and the highest review ratings given. From those results, it can be concluded that consumers who were satisfied by the book they had read wanted to express more of their personal opinion in their reviews, reflecting word-of-mouth scenarios where enthusiastic customers are often willing to provide more information about their experience and this will be reflected in their judgment of the product or service they have consumed.

5.2 Alignment with previous literature

The findings from this study confirmed those of a study by Mudambi and Schuff (Mudambi & Schuff, 2010), where a similar model was examined with a theoretical grounding involving divergence between experience and search goods. Regarding one aspect of their study related to helpfulness itself, it can be seen that Amazon UK reviews presented similar behavior in terms of helpfulness. However, although the length of the review in our case was significant, it had a zero coefficient. This indicates that, according to our findings in the case of Amazon UK, review length had no direct or indirect effect on helpfulness.

The results also provided more solid evidence for the discussed interplay between the readability and helpfulness scores attributed to an online review, evident in the previous study by Korfiatis et al. (Korfiatis et al., 2008). The present study constitutes an extension of this earlier article, by applying a more appropriate theoretical framework to the interplay between review helpfulness and text quality as assessed by readability tests.

Ghose and Ipeirotis (A. Ghose & Ipeirotis, 2010) developed readability metrics further, coupling them with extended data on reviewer characteristics, history and subjectivity scores computed by text-mining algorithms. Their econometric analysis provides interesting results in relation to the effect of review readability on online sales. In particular, they found that products (in that case digital cameras) which scored higher for readability also had higher sales. This study also provided a proxy for investigating the impact of readability as an independent measure for sales forecasting. Archak et al. (Archak et al., 2011), for instance, found evidence that qualitative aspects of the review text, estimated by externally imposed review semantics, may also play a role on the online sales of vertically differentiated products such as electronics.

6. Conclusions, limitations and further research

The most important result of our study has been to demonstrate that when a particular review is considered helpful by those who have already bought the product or will buy it in the future, this is related to the qualitative characteristics of the review justification as a piece of text. By employing readability formulas, we were able to analyze the reviews in our dataset and provide a set of results related to the helpfulness of a particular review as defined in this study.

In this study, we have focused on the content-specific characteristics of the review text. However, one of the limitations of this approach has been that we were unable to assess whether the review expressed a personal opinion about a product or a service or not. In fact, we know from the marketing literature that consumers tend to associate themselves with other consumers who express a more personal experience about the product which might influence consumers' product choice process (Bettman & Park, 1980). Another limitation of this study has been checking the actual reliability of the readability tests by cross validating whether the tests actually measure the readability of a review written on a website. This is because readability tests only take content factors into account, and do not address usability (e.g. the position of the text on the screen, etc.); thus, they can only provide an indication of whether the written text would be understandable for an average reader. Furthermore, it should be noted that the results we present in this study hold true for book items; reviews about other types of goods, such as search goods, might produce different effects on the relationship between HR and the qualitative characteristics of the review text. To this end this study could be extended further to examine whether the findings hold true for vertically differentiated products.

Another factor which could complement the discussed model would be the use of sentiment analysis. However, the readability measures which were used in the present study to operationalize our constructs do not consider lexical traits, which could indicate reviewer sentiments regarding the utility they received from consumption of the product or service (García-Barriocanal, Sicilia, & Korfiatis, 2010). This was beyond the scope of our study. In our model, we framed the extensiveness of the review length as expressiveness and we investigated its relationship to the review rating itself rather than to helpfulness.

This paper contributes to the ever-growing literature on the importance of online reviews as an advantage of online marketplaces over traditional markets (Ba & Pavlou, 2002, Hu et al., 2006, Pavlou et al., 2007, Stewart & Pavlou, 2002) . Codification of information related to products or services can actually help future buyers to evaluate the quality of an experience good (in our case,

books) by reading the judgments provided by other customers. We believe that the qualitative characteristics of online reviews are a rich source of information, contributing to our understanding of how consumers evaluate product information on an online marketplace. Based on this study, we aim to pursue this analysis further by incorporating consumers' cognitive characteristics as captured by their reviews, for example, the emotional expressions that are used in the text and which can be analyzed via specialized lexical resources (Bentivogli et al., 2004). Recent developments in computational linguistics related algorithms that asses this further, such as the Linguistic Inquiry and Word Count (Tausczik & Pennebaker, 2010), can also assist in researching this aspect in additional depth.

References

Archak, N., Ghose, A., & Ipeirotis, P. (2011). Deriving the pricing power of product features by mining consumer reviews. *Management Science*, mnsc. 1110.1370 v1.

Ba, S., & Pavlou, P. A. (2002). Evidence of the effect of trust building technology in electronic markets: Price premiums and buyer behavior. *MIS quarterly*, 243-268.

Chen, Y., Xie, J., & Hall, M. C. (2008). Online consumer review: Word-of-mouth as a new element of marketing communication mix. *Management Science*, *54*(3), 477-491.

Chevalier, J. A., & Mayzlin, D. (2006). The effect of word of mouth on sales: Online book reviews. *Journal of Marketing Research*, *43*(3), 345-354.

Clemons, E. K., Barnett, S., & Appadurai, A. (2007). The future of advertising and the value of social network websites: some preliminary examinations. *Proceedings of the ninth international conference on Electronic commerce* (p. 276). ACM.

Clemons, E. K., Gao, G. G., & Hitt, L. M. (2006). When Online Reviews Meet Hyper-differentiation: A Study of the Craft Beer Industry. *Journal of Management Information Systems*, *23*(2), 149-171.

Coleman, M., & Liau, T. L. (1975). A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, *60*(2), 283-284.

Crowley, A. E., & Hoyer, W. D. (1994). An integrative framework for understanding two-sided persuasion. *Journal of Consumer research*, *20*(4), 561-574.

Dellarocas, C. (2003). The Digitization of Word of Mouth: Promise and Challenges of Online Feedback Mechanisms. *Management Science*, *49*(10).

DuBay, W. H. (2004). The principles of readability. Costa Mesa. *CA: Impact Information*.

Eisend, M. (2006). Two-sided advertising: A meta-analysis. *International Journal of Research in Marketing*, *23*(2), 187-198.

Eroglu, S. A., Machleit, K. A., & Davis, L. M. (2003). Empirical testing of a model of online store atmospherics and shopper responses. *Psychology and Marketing*, *20*(2), 139-150.

Flesch, R. F. (1951). *How to Test Readability*. Harper.

Forman, C., Ghose, A., & Wiesenfeld, B. (2008). Examining the Relationship Between Reviews and Sales: The Role of Reviewer Identity Disclosure in Electronic Markets. *Information Systems Research*, *19*(3), 291-313.

García-Barriocanal, E., Sicilia, M.-A., & Korfiatis, N. (2010). Exploring hotel service quality experience indicators in user-generated content: a case using Tripadvisor data. *Proceedings of the 5th Mediteranian Conference on Information Systems (MCIS 2010)*. Presented at the MCIS 2010, Tel-Aviv, Israel: Association of Information Systems (AIS).

Gerdes, J., Stringam, B. B., & Brookshire, R. G. (2008). An integrative approach to assess qualitative and quantitative consumer feedback. *Electronic Commerce Research*, *8*(4), 217-234.

Ghose, A., & Ipeirotis, P. G. (2010). Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics. *IEEE Transactions on Knowledge and Data Engineering*, *23*(10).

Godes, D., Mayzlin, D., Chen, Y., Das, S., Dellarocas, C., Pfeiffer, B., Libai, B., et al. (2005). The Firm's Management of Social Interactions. *Marketing Letters*, *16*(3), 415-428.

Gunning, R. (1969). The Fog Index After Twenty Years. *Journal of Business Communication*, *6*(2), 3.

Hu, N., Pavlou, P. A., & Zhang, J. (2006). Can online reviews reveal a product's true quality?: empirical findings and analytical modeling of Online word-of-mouth communication. *Proceedings of the 7th ACM conference on Electronic commerce* (pp. 324-330).

Jiang, J., Gretzel, U., & Law, R. (2010). Do Negative Experiences Always Lead to Dissatisfaction?—Testing Attribution Theory in the Context of Online Travel Reviews. *Information and Communication Technologies in Tourism 2010*, 297-308.

Jiang, Z., & Benbasat, I. (2004). Virtual product experience: Effects of visual and functional control of products on perceived diagnosticity and flow in electronic shopping. *Journal of Management Information Systems*, *21*(3), 111-147.

Jiang, Z., & Benbasat, I. (2007). Investigating the influence of the functional mechanisms of online product presentations. *Information Systems Research*, *18*(4), 454-470.

Jurca, R., Garcin, F., Talwar, A., & Faltings, B. (2010). Reporting incentives and biases in online review forums. *ACM Transactions on the Web (TWEB)*, *4*(2), 1-27.

Kelman, H. C. (1958). Compliance, identification, and internalization: Three processes of attitude change. *The Journal of Conflict Resolution*, *2*(1), 51-60.

Kincaid, J. P., Fishburne Jr, R. P., Rogers, R. L., & Chissom, B. S. (1975). Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel.

Korfiatis, N., Rodríguez, D., & Sicilia, M.-A. (2008). The Impact of Readability on the

Usefulness of Online Product Reviews: A Case Study on an Online Bookstore. *Emerging Technologies and Information Systems for the Knowledge Society*, Lecture Notes in Computer Science (Vol. 5288, pp. 423-432). Springer Berlin / Heidelberg.

Kumar, N., & Benbasat, I. (2001). Shopping as experience and Web site as a social actor: Web interface design and para-social presence. *Proceedings of the Twenty-Second International Conference on Information Systems* (pp. 449-454).

Li, C., & Bernoff, J. (2008). *Groundswell: Winning in a world transformed by social technologies*. Harvard Business School Pr.

Moscovici, S. (1985). Social influence and conformity. *Handbook of Social Psychology: Special fields and applications*, 347.

Mudambi, S. M., & Schuff, D. (2010). What makes a helpful online review? A study of customer reviews on amazon. com. *MIS Quarterly*, *34*(1), 185-200.

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

Paasche-Orlow, M. K., Taylor, H. A., & Brancati, F. L. (2003). Readability Standards for Informed-Consent Forms as Compared with Actual Readability. *New England Journal of Medicine*, *348*(8), 721.

Park, D. H., & Kim, S. (2009). 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.

Parra, J. F., & Ruiz, S. (2009). Consideration sets in online shopping environments: the effects of search tool and information load. *Electronic Commerce Research and*

Applications, *8*(5), 252-262.

Pavlou, P. A., & Dimoka, A. (2006). The Nature and Role of Feedback Text Comments in Online Marketplaces: Implications for Trust Building, Price Premiums, and Seller Differentiation. *Information Systems Research*, *17*(4), 392-414.

Pavlou, P. A., Liang, H., & Xue, Y. (2007). Understanding and mitigating uncertainty in online exchange relationships: A principal-agent perspective. *Mis Quarterly*, *31*(1), 105-136.

Senecal, S., & Nantel, J. (2004). The influence of online product recommendations on consumers' online choices. *Journal of Retailing*, *80*(2), 159-169.

Shocker, A. D., Ben-Akiva, M., Boccara, B., & Nedungadi, P. (1991). Consideration set influences on consumer decision-making and choice: Issues, models, and suggestions. *Marketing letters*, *2*(3), 181-197.

Smith, M., & Taffler, R. (1992). Readability and Understandability: Different Measures ofthe Textual Complexity of Accounting Narrative. *Accounting, Auditing & Accountability Journal*, *5*(4).

Stewart, D. W., & Pavlou, P. A. (2002). From consumer response to active consumer: measuring the effectiveness of interactive media. *Journal of the Academy of Marketing Science*, *30*(4), 376-396.

Stringam, B. B., & Gerdes, J. (2010). An Analysis of Word-of-Mouse Ratings and Guest Comments of Online Hotel Distribution Sites. *Journal of Hospitality Marketing & Management*, *19*(7), 773-796.

Tausczik, Y. R., & Pennebaker, J. W. (2010). The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social*

Psychology, *29*(1), 24 -54.

Tsang, A. S. L., & Prendergast, G. (2009). Is a "star" worth a thousand words?: The interplay between product-review texts and rating valences. *European Journal of Marketing*, *43*(11/12), 1269-1280.

Van der Heijden, H., & Verhagen, T. (2004). Online store image: conceptual foundations and empirical measurement. *Information & Management*, *41*(5), 609-617.

Vessey, I., & Galletta, D. (1991). Cognitive fit: An empirical study of information acquisition. *Information Systems Research*, *2*(1), 63-84.

Wolfinbarger, M., & Gilly, M. C. (2001). Shopping Online for Freedom, Control, and Fun. *California Management Review, 43*(2).

Zakaluk, B. L., & Samuels, S. J. (1988). *Readability: Its Past, Present, and Future*. Newark: International Reading Association.

Zhang, Z., Ye, Q., Law, R., & Li, Y. (2010). The impact of e-word-of-mouth on the online popularity of restaurants: A comparison of consumer reviews and editor reviews. *International Journal of Hospitality Management*, *29*(4).

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.