“Highly Recommended!” The Content Characteristics and Perceived Usefulness of Online Consumer Reviews

Lotte M. Willemsen,
Peter C. Neijens,
Fred Bronner, and
Jan A. de Ridder

The Amsterdam School of Communication Research ASCoR University of Amsterdam

The aim of the present study was to gain a better understanding of the content characteristics that make online consumer reviews a useful source of consumer information. To this end, we content analyzed reviews of experience and search products posted on Amazon.com (N = 400). The insights derived from this content analysis were linked with the proportion of ‘useful’ votes that reviews received from fellow consumers. The results show that content characteristics are paramount to understanding the perceived usefulness of reviews. Specifically, argumentation (density and diversity) served as a significant predictor of perceived usefulness, as did review valence although this latter effect was contingent on the type of product (search or experience) being evaluated in reviews. The presence of expertise claims appeared to be weakly related to the perceived usefulness of reviews. The broader theoretical, methodological and practical implications of these findings are discussed.

Key words: Online Consumer Reviews, Electronic Word-of-Mouth Communication, Content Characteristics, Content Analysis.


With the emergence of consumer-generated media platforms, word-of-mouth conversations have migrated to the World Wide Web (Brown, Broderick, & Lee, 2007), creating a wealth of product information that is often articulated in the form of online consumer reviews (Schindler & Bickart, 2005). These reviews provide product evaluations from the perspective of the customer, and have a strong influence on consumers’ product and brand attitudes and purchase behavior (Chevelier & Mayzlin, 2006; D.-H. Park & Kim, 2008; Senecal & Nantel, 2004), even more so than marketer-generated information (Chiou & Cheng, 2003). The persuasive impact of online consumer reviews, as well as of other forms of word-of-mouth, is often attributed to the perceived non-commercial nature of their authors. Consumers are believed to have no vested interest in recommending a product or brand, and their implied independence renders reviews more credible and consequently more useful than marketer-generated information (Bickart & Schindler, 2001; Ha, 2002; Herr, Kardes, & Kim, 1991).

As reviews gain in popularity, it becomes harder for consumers to find their way in the wealth of reviews and to assess the usefulness of the information offered (D.-H. Park & Lee, 2008). To circumvent
the problem of information overload, many review websites have invested in peer-rating systems that enable consumers to vote on whether they found a review useful in their purchase decision-making process. These votes serve as an indicator of review diagnosticity, and are used as a signaling cue to users to filter relevant opinions more efficiently (Ghose & Ipeirotis, 2008; Mudambi & Schuff, 2010).

Variations in the proportion of ‘useful’ votes provide evidence that ‘all reviews are not created equal’ (Godes & Mayzlin, 2004; D.-H. Park, Lee, & Han, 2007) and, hence, that all reviews are not evaluated as equal. Consumers do not follow a structured format when posting their product evaluations on the web (Park & Kim, 2008; Pollach, 2008). As a consequence, reviews range from simple recommendations that are accompanied by extremely positive or negative statements, to nuanced product evaluations that are supported by extensive reasoning. However, hardly any research has been conducted in order to catalogue differences in the content of reviews, or study the impact of such differences on the perceived usefulness of reviews (Mudambi & Schuff, 2010).

To fill this research gap, our research aims at gaining a better understanding of the content characteristics that make online consumer reviews a useful source of information. More specifically, we seek to understand how three types of content characteristics—that is, expertise claims, review valence and argumentation style—affect the perceived usefulness of reviews. In addressing these aims, we perform a systematic content analysis of consumer reviews and the proportion of useful votes that reviews received from fellow consumers via peer rating mechanisms. This research endeavor extends research on online consumer reviews in two ways.

First, this study responds to the need to perform content analyses to gain more insight into the composition of reviews (Mazzarol, Sweeney, & Soutar, 2007; Schindler & Bickart, 2005). Content analyses have heretofore been missing due to a lack of proper measurement tools to process the linguistic complexity of online reviews (Godes & Mayzlin, 2004; Mudambi & Schuff, 2010; Godes et al., 2005). Mixtures of closely linked positive and negative statements, domain-specific language (e.g., technical terms to describe product attributes), anecdotal information and the lack of grammar and structure have challenged efforts to document the content of reviews, manually and automatically (Ganu, Elhadad & Marian, 2009). This study employs a relational content analysis that was especially developed to unravel complex evaluative discourses: the Network analysis of Evaluative Texts (hereafter: NET-method, see Van Cuijlenburg, Kleinnijenhuis, & De Ridder, 1988). The results deriving from this analysis represent a first step towards a better understanding of the nature of reviews.

Second, this study uses real data to unpack the content characteristics that drive people to respond and value online consumer reviews. By using measures that let people speak for themselves in an unsolicited manner—i.e. the proportion of useful votes—and linking these with the results from the content analysis, we follow up on Schindler and Bickart’s (2005) call to “take advantage of the frozen chunks of word-of-mouth exchanges saved on the Internet to more effectively study what makes a persuasive message” (p. 58).

The Content Characteristics and Perceived Usefulness of Reviews

Online consumer reviews contain open-ended comments and ratings (D.-H. Park & Kim, 2008). Open-ended comments display reviewers’ assessments of the positive and/or negative qualities of a product as voiced in the textual content of reviews. Ratings are numeric summary statistics, often prominently shown in the form of five-point star recommendations at the surface level of the review, and encapsulate reviewers’ general assessments of the product. In addition to ratings that reflect reviewers’ products assessments, most review sites nowadays also publish ratings that reflect users’ review assessments. The perceived usefulness of a review serves as the primary currency to gauge how users evaluate a review. Expresssed by an annotation such as ‘10 out of 12 people found the following review useful’, the
perceived usefulness of reviews appears along with the product ratings at the surface level of the review (see Figure 1, panel A).

The perceived usefulness of a review has been found to be a significant predictor of consumers’ intent to comply with a review (Cheung, Lee, & Rabjohn, 2008). Interpreting perceived usefulness as “a measure of perceived value in the purchase decision-making process” (Mudambi & Schuff, 2010, p. 186), scholarly research has recently started to explore the factors driving the perceived usefulness of reviews. This pioneering work shows that insight into the composition of reviews is imperative in understanding the effects of reviews on consumer judgment, as consumers seem to react differently to different types of reviews. For example, by linking useful votes to the product rating of a review, several studies found that clearly negative or positive product ratings (i.e., 1- and 5-star ratings) are perceived as more useful than moderate ratings (i.e., 3-star ratings, see Danescu-Niculescu-Mizil, Kossinets, Kleinberg, & Lee, 2009; Forman, Ghose, & Wiesenfeld, 2008). Others found that the polarity of product ratings contribute to the perceived usefulness of reviews, such that negative reviews have more impact on consumer judgment than positive reviews (Basuroy, Chatterjee, & Ravid, 2003; Chevelier & Mayzlin, 2006; Sen & Lerman, 2007).

Although these studies have broadened our knowledge with regard to the perceived usefulness of product ratings, there is more to a review than its rating that makes it a useful decision-aid, that is: its textual content. Recent research suggests that the content of eBay and YouTube comments provides a nuanced view of the positive and/or negative qualities of the object under review (e.g., retailers, film clips), and contributes more to the perceived usefulness of eBay and YouTube comments than numeric ratings (Lu, Zhai, & Sundaresan, 2009; Siersdorfer, Chelaru, Nejdl & Pedro, 2010). However, no study has examined the contribution of positively and negatively valenced statements to the perceived usefulness of reviews above and beyond numeric star ratings. Also content characteristics pertinent to the persuasive impact of message content, such as argumentation style and the presence of expertise claims have garnered little consideration in the literature although such characteristics are understood to be significant (Cialdini, 2001; McGuire, 1985; Petty & Cacioppo, 1984). They offer explanation and context to product ratings, and as such may be important drivers of a review’s perceived usefulness (cf. Mudambi & Schuff, 2010).

Against this background, we expect to find that the open-ended comments of reviews, and in particular three content characteristics within open-ended comments of reviews—i.e., expertise claims, review valence and argumentation—contribute to the perceived usefulness of reviews.

**Hypotheses Development**

**Expertise claims**

When making purchase decisions, consumers are inclined to seek and follow the advice of expert sources (e.g., Bansal & Voyer, 2000; Gilly, Graham, Wolfinbarger, & Yale, 1998) as they are believed to provide information that is based, more than non-expert ones, on accurate representations of reality (Hovland, Janis, & Kelley, 1953). Information that is attributed to an expert is found to be more trustworthy and useful, and to have more influence on the brand attitudes, purchase intentions and purchase behaviors of consumers (Harmon & Coney, 1982; Lascu, Bearden, & Rose, 1995).

The degree to which a source is considered an ‘expert’ is determined by evaluating the knowledge and competence that a source holds regarding the topic of interest (Gotlieb & Sarel, 1991). However, in online settings it is difficult to make such an evaluation given the limited availability of personal information (Cheung et al., 2008; Schindler & Bickart, 2005). Impression formation takes place in a reduced and altered cues environment in which the author’s attributes and background cannot
be verified. Evaluations of expertise must therefore be based on “the text-based resource exchange provided by actors” (Brown et al., 2007, p. 7), that is, based on reviewer’s self-claims.

If reviewers claim to have expert knowledge regarding the product under consideration, which reviewers frequently do according to a content analysis by Pollach (2008), their evaluations are likely to be perceived as useful. Indirect support for such an expectation has been provided by Eastin (2001), who studied the effects of source expertise on the credibility of online health information. He found that people use expertise to evaluate the credibility of unfamiliar information. A message authored by a person who identified him-/herself as an expert was deemed more credible than a message authored by a person who identified him-/herself as a layperson. This finding has been explained by the literature through the operation of the ‘authority’ heuristic (cf. Hu & Sundar, 2010; Tan, Swee, Lim, Detenber, & Alsagoff, 2008): a cognitive decision rule (‘expert statements are true’) established through prior experience that teaches that experts are a valid source of information due to their authority on a subject. This heuristic inference steers people’s judgment of a message whenever a source’s expertise becomes salient, such that it will be positively evaluated irrespective of the kind of information offered (e.g. Chaiken, Liberman, & Eagly, 1989). Hence, we hypothesized the following:

H1. The higher a reviewer’s claimed expertise, the more useful a review is perceived to be.

Review Valence
The literature has demonstrated a positive relationship between review valence and consumer behavior (Clemons, Gao, & Hitt, 2006; Dellarocas, Awad Farag, & Zhang, 2004; D.-H. Park et al., 2007). The more positive (negative) the valence of a review, the more (less) likely people are to purchase the reviewed product. However, some studies have shown that consumers give greater weight to negative than positive reviews. For example, Basuroy and colleagues (2003) demonstrated that negative reviews hurt a movie’s box office performance more than positive reviews contribute to its performance. Similarly, in a study that examined the effect of consumer reviews on relative sales of books at Amazon.com and Barnesandnoble.com, Chevalier and Mayzlin (2006) found that reviews with 1-star ratings have a stronger effect on sales figures than reviews with 5-star ratings. These findings are indicative of a negativity bias, whereby negative information has a systematically stronger impact on judgment and choice than objectively equivalent positive information (Skowronski & Carlston, 1989).

The negativity bias has been explained by a variety of mechanisms, including that negative information is more novel and attention-grabbing and has greater salience (e.g., Fiske, 1980; Pratto & John, 1991). The most accepted explanation for the asymmetry in valence evaluations is offered by the category diagnosticity theory (Ahluwalia, 2002; Herr et al., 1991; Skowronski & Carlston, 1989), which holds that negative information is more diagnostic and useful than positive information when classifying products according to quality and performance. This is because negative product attributes are thought to be characteristic primarily of low quality products, while positive product attributes are thought to be characteristic of both low and high quality products.

Scholars have recently postulated that negative information may not be equally diagnostic for all products because of differences in the pre-purchase performance veracity (C. Park & Lee, 2009; Xia & Bechwati, 2008). In terms of pre-purchase performance veracity, products can be classified as either search products or experience products (Girard, Silverblatt, & Korgaonkar, 2002; Nelson, 1970). Search products, such as electronics, are products that can be accurately evaluated prior to purchase because they are characterized by concrete and functional attributes for which valid information can be obtained before product use. Experience goods, such as recreational services, are dominated by
intangible attributes that cannot be known until purchase, and for which performance evaluations can be verified only by (sensory) experience or consumption.

When the attributes of a product are intangible or not immediately verifiable—which is the case with experience products—there is a greater chance of making an incorrect decision. Accuracy concerns steer consumers to adopt a risk-aversive outlook when reading reviews of experience products, thereby leading to a higher perceived diagnosticity and hence greater scrutiny and evaluation of negative information (Ahluwalia, 2002). In contrast, when a review evaluates a search product for which attribute information is verifiable and the chance of loss is not very likely, people may be less inclined to seek and value negative information. Accordingly, we expected review valence and product type to interact:

H2. Negatively valenced reviews induce a higher perceived usefulness than positively valenced reviews. This effect is more pronounced for experience products than for search products.

**Argumentation**

Reviews contain not only valenced statements in the open-ended comments of reviews (D.-H. Park & Kim, 2008), but also arguments to support those valenced statements. Whereas research has delved into the valence of reviews, hardly any research has been conducted on the presence of arguments in reviews. This is surprising, since the effect of argumentation has been well-established in a myriad of empirical studies in communication science. These studies show that the proportion of arguments in messages is positively related to people’s intent to comply with those messages (e.g. Petty & Cacioppo, 1984; Price, Nir, & Cappella, 2006; Raju, Unnava, & Montgomery, 2009). As explained by O’Keefe (1998), explicitly articulating the arguments upon which an opinion is based “opens the advocated view for critical scrutiny” (p. 61). The mere presence of arguments consequently leads people to have more confidence in a communicator and to find his/her judgment more persuasive.

In the vein of computer-mediated communication, online discussants may be especially apt to judge information based on the rigor of arguments, as they are not able to rely on social cues—such as gestures, intonation and facial expressions—that serve to validate information in face-to-face settings (Walther, 1996). In a study by Price, Nir and Cappella (2006) that examined 60 online group discussions about the tax plans offered by rival presidential candidates George W. Bush and Al Gore, it was indeed found that the mere presence of arguments makes online messages more persuasive. Although the focus of this study was on political behavior rather than on consumer behavior, and on synchronous rather than asynchronous communication, similar forces can be expected for online consumer reviews. A qualitative study by Schindler and Bickart (2005) indicates that since reviewers are anonymous on the Internet, people will not easily accept or believe a review if it does not provide sufficient information on the arguments used when making claims about or evaluating a product or service. We therefore hypothesized that reviews are more valuable if they offer more arguments to back up their valenced statements (hereafter ‘argument density’):

H3. The higher the argument density, the more useful a review is perceived to be.

In addition to the density of argumentation, the content of the argumentation may play a powerful role in what consumers gain from reviews. For instance, a substantive finding in the persuasion literature is that texts that present both sides of an argument are generally more persuasive than other forms of texts (Eisend, 2007). Specifically, people are more likely to align their attitudes and brand and product preferences with the content of a message after reading the pros and cons of a position (i.e. two-sided argumentation) than after reading the pros or cons alone (i.e. one-sided argumentation). This effect has been explained in the context of attribution theory (e.g. Easley, Bearden, & Teel, 1995;
Eisend, 2007; Kamins, 1989; Kelley, 1972), which suggests that recipients’ views of why someone is sharing information influences how the information is received. The inclusion of negative information along with positive information serves as a validation cue that the text derives from a consumer who has authentic first-hand experience, and not from a commercial endorser who has an interest in recommending a product. Such inferences are important because of consumers’ concern that the anonymous nature of online reviews encourages commercial endorsers to write product evaluations under false consumer identities in order to promote products and brands (Schindler & Bickart, 2005). Consumers will discredit a review if they suspect that the source is not telling the truth, and that the recommendation is not based on an accurate evaluation of the product under consideration (Crowley & Hoyer, 1994; Pechmann, 1992). Following this line of reasoning, we hypothesized that reviews are perceived as more useful if they rely on a large diversity of positive and negative arguments (hereafter ‘argument diversity’):

**H4.** The higher the argument diversity, the more useful the review is perceived to be.

**Method**

Testing hypotheses 1–4 required a systematic content analysis of a varied sample of reviews on both search and experience products. With this requirement in mind, it was decided to content analyze reviews from Amazon.com. Amazon.com is the largest online retailer (in terms of international revenue and website visits, Chevelier & Mayzlin, 2006), that not only provides consumers with the opportunity to order goods from a wide range of product categories, but also to read and post reviews on those goods. Particularly important for the present study, is that Amazon.com also allows consumers1 to cast a vote on whether posted reviews were useful to them in the purchase decision-making process. As such, Amazon.com enabled us to analyze the relationship between content characteristics and perceived usefulness of online consumer reviews.

**Product Selection**

Before collecting the reviews from Amazon.com, we first had to identify products most appropriate to represent the product category variable. This was done in two steps. In the first step we selected a list of products offered on Amazon.com that, according to the definitions of Nelson (1970), represented search or experience products. In line with these definitions (see p. 9), four products were identified as search products (i.e., a digital camera, a laser printer, a DVD player and a food processor) and five products as experience products (i.e., sunscreen, an espresso machine, running shoes, shaving equipment and diet pills).

As products can be categorized along a continuum from pure search to pure experience, we performed a pilot test in step 2 to ascertain which of these Amazon.com products is most representative of each end of the continuum. On the basis of Krishnan and Hartline (2001, cf. Bronner & De Hoog, 2009), undergraduate students \((n = 50)\) were asked to indicate their ability to judge the performance of each product (a) before use and (b) after use, using a seven-point scale ranging from ‘Not at all’ to ‘Very well’. The results as presented in Table 1 showed that sunscreen and running shoes can be viewed as experience products, given the relatively low mean scores on the ‘before use’ scale \((M_{\text{sunscreen}} = 3.2; M_{\text{running shoes}} = 3.8)\) and the high mean scores on the ‘after use’ scale \((M_{\text{sunscreen}} = 5.7, M_{\text{running shoes}} = 6.2)\). Digital cameras and DVD players, in contrast, can be considered search products as they received high mean scores on both the ‘before use’ \((M_{\text{cameras}} = 4.6, M_{\text{dvd players}} = 4.8)\) and the ‘after use’ scale \((M_{\text{camera}} = 6.2, M_{\text{dvd players}} = 6.4)\). Furthermore, evaluation differences between
Table 1 Pretest to Determine Respondents’ Ability to Judge the Performance of Products Before and After Purchase

<table>
<thead>
<tr>
<th></th>
<th>Before Purchase</th>
<th></th>
<th>After Purchase</th>
<th></th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>SD</td>
<td>M</td>
<td>SD</td>
<td>M</td>
</tr>
<tr>
<td>Food processor</td>
<td>4.18</td>
<td>1.56</td>
<td>6.30</td>
<td>0.91</td>
<td>2.12</td>
</tr>
<tr>
<td>Sunscreen</td>
<td>3.18</td>
<td>1.55</td>
<td>5.68</td>
<td>1.32</td>
<td>2.50</td>
</tr>
<tr>
<td>Espresso machine</td>
<td>4.16</td>
<td>1.41</td>
<td>6.30</td>
<td>0.86</td>
<td>2.14</td>
</tr>
<tr>
<td>Digital camera</td>
<td>4.58</td>
<td>1.50</td>
<td>6.20</td>
<td>1.14</td>
<td>1.62</td>
</tr>
<tr>
<td>Running Shoes</td>
<td>3.76</td>
<td>1.71</td>
<td>6.24</td>
<td>1.06</td>
<td>2.48</td>
</tr>
<tr>
<td>DVD player</td>
<td>4.78</td>
<td>1.46</td>
<td>6.42</td>
<td>0.78</td>
<td>1.64</td>
</tr>
<tr>
<td>Printer</td>
<td>4.36</td>
<td>1.57</td>
<td>6.34</td>
<td>0.80</td>
<td>1.98</td>
</tr>
<tr>
<td>Diet pills</td>
<td>1.88</td>
<td>1.29</td>
<td>4.94</td>
<td>1.59</td>
<td>2.05</td>
</tr>
<tr>
<td>Shaving equipment</td>
<td>3.78</td>
<td>1.56</td>
<td>6.18</td>
<td>1.06</td>
<td>2.40</td>
</tr>
</tbody>
</table>

Note. Mean values on a 7-point scale, where 1 indicated “not at all” and 7 indicated “very well”.

the before and after use scales were significantly higher for experience than for search products, $F(3,196) = 4.91, p < .01$, indicating a clear difference in pre-purchase performance veracity. The results correspond with other studies that sought differences between the experience and search attributes of products (cf. Archibald, Haulman, & Moody, 1983; D.-H. Park & Han, 2008).

Sample Selection

Based on the results of the pilot test, reviews that had been posted between 2005 and 2009 and had at least one useful vote were extracted from Amazon.com (Mudambi & Schuff, 2010). The population comprised 42,700 reviews covering 38,745 reviews of cameras, 2,497 of DVD players, 1,032 of running shoes and 426 of sunscreen. To ensure equal group sizes for experience and search products, the uploaded reviews were subjected to a stratified random sampling method, with product type as stratum. This procedure resulted in a sample of 400 reviews equally distributed over the two experience and search product categories.

Content Analysis

To unravel the valence of product evaluations in reviews, as well as the arguments used to support those evaluations, this study employed the NET method (Van Cuilenburg et al., 1988); a relational content analysis (Popping, 2000) that enables one to extract from a given text a network of objects (e.g. actors, values, issues). Although never applied to online reviews or any other form of word-of-mouth, the NET method was opted as it has proven to be a useful means to analyze valence and argumentative structures in evaluative texts. Moreover, the NET method enables one to code actors and their personal characteristics, which was needed to tap reviewer’s claimed level of expertise (e.g. Kleinnijenhuis, De Ridder, & Rietberg, 1997).

‘Subject/Predicate/Object’ Triples as Core Phrases. The NET method divides a text into core phrases (Kleinnijenhuis et al., 1997): statements that describe the relations between objects in the form of triples. These triples consist of a predicate with a positive, neutral or negative meaning to indicate the degree of association/dissociation of a subject with an object, ranging from $-1$ (maximal dissociation) to $+1$ (maximal association). For example, if a reviewer states that the ‘camera is very easy to use’,
(2) This camera is very easy to use, which is especially important to me since I don’t know anything about cameras (3) The only thing is that this camera is quite heavy, and therefore a burden to carry around.

Was this review useful to you?  Yes  No

<table>
<thead>
<tr>
<th>Sentence</th>
<th>Subject</th>
<th>Predicate</th>
<th>Association</th>
<th>Object</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1)</td>
<td>Product</td>
<td>Recommended</td>
<td>+ 1</td>
<td>Ideal</td>
</tr>
<tr>
<td>(2)</td>
<td>Product</td>
<td>easy to use</td>
<td>+ 1</td>
<td>Ease of use</td>
</tr>
<tr>
<td>(2)</td>
<td>Easy of use</td>
<td>important</td>
<td>+ 1</td>
<td>Ideal</td>
</tr>
<tr>
<td>(2)</td>
<td>Reviewer</td>
<td>Don’t know</td>
<td>– 1</td>
<td>expert</td>
</tr>
<tr>
<td>(3)</td>
<td>Product</td>
<td>heavy</td>
<td>+ 1</td>
<td>heavy</td>
</tr>
<tr>
<td>(3)</td>
<td>heavy</td>
<td>burden</td>
<td>– 1</td>
<td>Ideal</td>
</tr>
</tbody>
</table>

Figure 1

the reviewer associates the product with the attribute ‘ease of use’ (i.e. product/ +1/ease of use). In a similar vein, if a reviewer claims that s/he has no knowledge about a camera, the reviewer dissociates her-/himself from the attribute ‘expertise’ (i.e. reviewer/-1/expert). In the case of purely evaluative statements, a subject is associated/ dissociation with a special object called ‘Ideal’, which represents a positive evaluation of the object under consideration. Thus, a sentence like ‘this product is highly recommendable’ is coded as a core phrase in which the product is associated with the positive Ideal by calling it recommendable (i.e. product/+1/Ideal) (see Figure 1 for more example codings).

Direct and Indirect Core Phrases. The object of one core phrase may be the subject of another core phrase (see Figure 1, panel B). Hence, by coding single statements in the form of ‘subject/predicate/object’, the NET method provides quantitative measures of the whole network of relations between objects (see Figure 1, panel C). As a feature of networks, these relations can be either direct or indirect (when interconnected). By combining direct and indirect relations via summation
and multiplication (De Ridder, 1994), one can construct the argumentative structure of a text as well as its valence. The rationale underlying multiplication is that of evaluative transitivity (Van Cuilenburg et al., 1988): if product X scores well on ‘ease of use’ and ‘ease of use’ is evaluated as a desirable attribute of the product, then this implies that product X will also be evaluated as desirable. The direction of such indirect relations gauges the valence of arguments, since interconnected relations amount to chain arguments: claims that evaluate an object in terms of its (un)favorable consequences or (dis)advantages (cf. Perelman & Olbrechts-Tyteca, 1969).

Coding Procedure
Reviews were coded sentence by sentence in accordance with the NET method. Reviews were content analyzed only if they contained core phrases that emphasized relations between the reviewer, product, reviewer/product attribute(s) and the evaluative object 'Ideal'. Reviews that did not meet this criterion (e.g. reviews that evaluated Amazon.com rather than a product offered on Amazon.com) were excluded from the analysis, resulting in a final sample size of 388 reviews.

The coding was supported by a semi-automated computer program, iNET. Six coders were trained in using the computer program and coding instructions. Throughout the coding period, each coder analyzed about 15% of the sample. An additional 10% was analyzed by two or more coders in order to determine intercoder reliability. Using the F1-score (a measure that computes the similarity of the networks extracted from the same texts by different coders) the overall agreement was .77, which provided a good level of intercoder reliability based on the criteria of Landis and Koch (1977). More precisely, the F1 score was .73 for reviews of search products, and .81 for reviews of experience products.

Measures
Valence. The valence of reviews was operationalized as the weighted mean value of core phrases with the product under review as subject and Ideal as object. A statement was coded as a core phrase with the product as subject and Ideal as the object if the reviewer considered the product good, essential, virtuous, praiseworthy or capable. Weighted mean scores range from —1 (negative valence) to +1 (positive valence).

Argument Density. As indirect valenced statements are considered arguments (Kleinnijenhuis et al., 1997), we calculated argument density as the proportion of indirect core phrases with Ideal as object (e.g., statements where a product is evaluated based of its (dis)advantages in terms of weight or ease of use, see sentence 2 and 3 in Figure 1) as opposed to the total number of direct and indirect core phrases with Ideal as object (e.g., all evaluative statements about the product, see sentence 1-3 in Figure 1). The resulting score is expressed in percentages ranging from 1 to 100 and presents the degree to which evaluative statements are substantiated by arguments.

Argument Diversity. To gain insight into the diversity of positive and negative arguments, we calculated the variance of the values of indirect valenced core phrases (see De Ridder, 1994, for the calculation operation). Scores range from 0 (low diversity) to +1 (high diversity).

Expertise Claims. Reviewers’ expertise claims were operationalized as the weighted mean of the values of direct core phrases in which the reviewer (dis)associates him-/herself with the trait ‘expertise’. Expertise was defined as all statements in a review that emphasize the reviewer’s product or product class knowledge that is derived from experience, study or training (Friedman & Friedman, 1979). Claimed expertise scores range from —1 (no expertise) to +1 (high expertise).
**Perceived Usefulness**

We collected the useful votes and total votes given above posted reviews in the form: “[number of useful votes] out of [number of members who voted] found the following review useful” (see Figure 1). By calculating the fraction of useful votes among the total votes, useful votes were translated into percentages ranging from 1 to 100 that indicate the ‘perceived usefulness of a review’ (Forman et al., 2008; Mudambi & Schuff, 2010).

**Control Variables.** To control for the effects of possible confounding variables, we collected several product and reviewer characteristics as mentioned at the surface level of the review as these were found to affect review effects in prior research (Chevelier & Mayzlin, 2006; Ghose & Ipeirotis, 2008). Product-related controls included the price of the product and the star rating (number of stars assigned to the product by the reviewer). Reviewer-related controls included the reviewer’s disclosure of real name (0 = no disclosure of real name; 1 = disclosure of real name) and place of residence (0 = no disclosure of location; 1 = disclosure of location), and his/her reputation as a top-1000 reviewer (0 = no top-1000 reviewer; 1 = top-1000 reviewer). Finally, we measured certain message-related factors, such as the length of the message (i.e. number of words) and the elapsed date (i.e. number of days since the posting of the review).

**Results**

**Search versus Experience Products**

The present study expects product category to play an important role in determining the relationship between content characteristics and perceived usefulness. We therefore collected reviews of four types of products that, according to the pilot test, could be assigned to either the search product category or the experience product category. To test whether reviews of DVD -players and digital cameras (search products) and running shoes and sunscreen (experience products) showed differences in content characteristics, we performed a series of t-tests. The results yielded no significant differences for DVD players and digital cameras in terms of review valence ($t(198) = 1.13, p = .26$), expertise claims ($t(198) = -.19, p = .85$), argumentation density ($t(198) = 1.21, p = .23$) and argument diversity ($t(198) = 1.43, p = .15$). Hence, reviews of both products were taken together to represent the search product category.

When comparing content characteristics for running shoes and sunscreen, no significant differences were found regarding review valence ($t(186) = -.13, p = .89$), expertise claims ($t(196) = 1.52, p = .13$) and argument diversity ($t(186) = .55, p = .59$). An exception is the significant difference for argumentation density ($t(169) = -3.57, p < .001$). Argumentation was significantly more dense in reviews of sunscreen ($M = 81.58, SD = 19.9$) than in reviews of running shoes ($M = 68.02, SD = 29.06$). However, since this is the only message characteristic on which reviews of running shoes and sunscreen differed, it was decided to take these reviews together to represent the experience product category. Summary statistics of content characteristics for reviews are given in Table 2.

**Regression Analyses of Content characteristics on Perceived Usefulness**

A series of ordinary least squares (OLS) regressions were performed to assess whether content characteristics add to the perceived usefulness of reviews above and beyond general characteristics as shown at the surface level of reviews. In these analyses, purchase price, star rating, elapsed date, length, top-reviewer status and self-disclosure (name and location) as control variables in the baseline model (Model 0). Model 1 included expertise claims in addition to the aforementioned control variables in
Table 2 Descriptive Statistics for Reviews

<table>
<thead>
<tr>
<th></th>
<th>M</th>
<th>SD</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Useful Votes</td>
<td>6.75</td>
<td>18.74</td>
<td>1</td>
<td>303</td>
</tr>
<tr>
<td>Perceived Usefulness (%)</td>
<td>77.15</td>
<td>33.32</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>Star Rating Review</td>
<td>4.00</td>
<td>0.67</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>Length Review (no. words)</td>
<td>115.43</td>
<td>97.77</td>
<td>18</td>
<td>797</td>
</tr>
<tr>
<td>Expertise claims</td>
<td>0.25</td>
<td>0.50</td>
<td>-1</td>
<td>1</td>
</tr>
<tr>
<td>Review Valence</td>
<td>0.41</td>
<td>0.67</td>
<td>-1</td>
<td>1</td>
</tr>
<tr>
<td>Argument Density (%)</td>
<td>74.25</td>
<td>25.21</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>Argument Diversity</td>
<td>0.07</td>
<td>0.21</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

order to test H1. Valence and product type were added in the regression equation in Model 2. This latter model also included the interaction term between product type and review valence to test H2. Before computing this interaction term, review valence was centred to minimize potentially problematic multicollinearity (Aiken & West, 1991). Argument density and diversity were finally entered in Model 3 to test H3 and H4. Table 3 presents the results of this regression analyses and reports the contribution of each of the independent variables toward perceived usefulness.

Model 0: Control Variables. At the baseline model, four surface characteristics appeared to be of significance. Purchase price had a negative and significant effect on perceived usefulness ($\beta = -0.14, p < .01$). Review length ($\beta = 0.24, p < .001$), star rating ($\beta = 0.22, p < .001$) and location disclosure ($\beta = 0.10, p < .05$) all demonstrated a significant positive effect on perceived usefulness, which together with purchase price accounted for 13.2% of the variance in perceived usefulness, $F(6, 351) = 6.09, p < .001$. Thus, longer reviews that assign more stars to a product and disclose the reviewer’s place of residence are considered more useful, as are reviews of relatively low-priced products.

Model 1: Expertise Claims. The inclusion of expertise claims in Model 1 resulted in a marginally significant change in the total amount of variance explained ($\Delta R^2 = .009, F_{change}(1, 350) = 3.70, p < .055$). The relationship between expertise claims and perceived review usefulness approached significance ($\beta = .10, p = .055$), thereby providing directional support for H1 stating a positive relationship between expertise claims and the perceived usefulness of a review.

Model 2: Review Valence and Product Type. In model 2, we added review valence, product type (search vs. experience, dummy coded) and the interaction term between product type and review valence. Overall, the three variables explained an additional 2.0% of the variance in perceived usefulness, $F_{change}(3, 347) = 2.74, p < .05$. Results showed that review valence had a marginally positive main effect on perceived usefulness ($\beta = .14, p = .08$). In addition, and consistent with H2, this effect was qualified by a highly significant interaction between review valence and product type ($\beta = -.18, p < .01$). The interaction effect (see Figure 2) showed that the negativity effect was prevalent only for experience products. In fact, simple slopes analysis, using the procedures described by Aiken and West (1991), demonstrated that review valence was negatively related to the perceived usefulness of reviews discussing experience products ($\beta = -.15, p < .05$, one-tailed), and positively related to the perceived usefulness of reviews discussing search products ($\beta = .14, p < .05$, one-tailed).

Model 3: Argument Density and Diversity. The inclusion of argument density and argument diversity made a significant contribution, adding another 3.3% of the variance in perceived usefulness,
Table 3 Hierarchical Multiple Regression Analysis Predicting the Perceived Usefulness of Online Consumer Reviews From Content Characteristics

<table>
<thead>
<tr>
<th>Model 0</th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Purchase Price</td>
<td>−.136**</td>
<td>−.133**</td>
<td>−.129*</td>
</tr>
<tr>
<td>Star Rating Review</td>
<td>.219***</td>
<td>.211***</td>
<td>.186**</td>
</tr>
<tr>
<td>Elapsed date</td>
<td>.004</td>
<td>.002</td>
<td>.014</td>
</tr>
<tr>
<td>Length of Review</td>
<td>.240***</td>
<td>.217***</td>
<td>.232***</td>
</tr>
<tr>
<td>Disclosure Real Name</td>
<td>.037</td>
<td>.031</td>
<td>.041</td>
</tr>
<tr>
<td>Disclosure Location</td>
<td>.099*</td>
<td>.217†</td>
<td>.085†</td>
</tr>
<tr>
<td>Top 1000 Reviewer</td>
<td>−.039</td>
<td>−.031</td>
<td>−.018</td>
</tr>
<tr>
<td>Expertise claims</td>
<td>—</td>
<td>−.100†</td>
<td>.086†</td>
</tr>
<tr>
<td>Review Valence</td>
<td>—</td>
<td>—</td>
<td>.141†</td>
</tr>
<tr>
<td>Product Type</td>
<td>—</td>
<td>—</td>
<td>.046</td>
</tr>
<tr>
<td>Review valence x Product Type</td>
<td>—</td>
<td>—</td>
<td>−.180**</td>
</tr>
<tr>
<td>Argumentation Density</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Argument Diversity</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>R²</td>
<td>.122</td>
<td>.131</td>
<td>.151</td>
</tr>
<tr>
<td>R² change</td>
<td>—</td>
<td>.009†</td>
<td>.020*</td>
</tr>
<tr>
<td>F-value</td>
<td>6.975***</td>
<td>6.614***</td>
<td>5.622***</td>
</tr>
</tbody>
</table>

Note: Coefficients are standardized regression coefficients (betas). †<.10; *p < .05; **p < .01; ***p < .001

Figure 2 Perceived Usefulness by Review Valence and Product Category
$F_{\text{change}}(2, 345) = 6.88, p < .001$. As predicted, argument density appeared to be a significant predictor of perceived usefulness ($\beta = .15, p < .01$). In line with H3, reviews are evaluated as more useful when product evaluations are supported by a higher number of arguments. Support was also found for H4. Model 3 in Table 3 revealed a positive relationship between argument diversity and perceived usefulness ($\beta = .14, p < .05$), indicating that negative arguments along with positive arguments contribute to a higher perceived usefulness.

**Discussion**

The aim of the present study was to gain a better understanding of the content characteristics that make online consumer reviews a useful source of information. To address this aim, we performed a content analysis of reviews discussing experience and search products offered by Amazon.com and the usefulness scores that reviews received from fellow consumers. The results indicate, after controlling for a variety of characteristics shown at the surface level of reviews (e.g. star rating, characteristics of reviewer, purchase price of product), that differences in the perceived usefulness of reviews are related to differences in the content of reviews. We identify this finding as an indication that as ‘all reviews are not created equal’, all reviews are not evaluated as equal.

This becomes evident in the finding that the relation between review valence and perceived review usefulness differs for experience and search products. Prior research reported a negativity effect for the effects of review valence, showing that negative information has a stronger impact on judgment and choice than objectively equivalent positive information (Godes & Mayzlin, 2004; D.-H. Park et al., 2007). In the research reported here, this negativity effect was present only for experience products: negatively valenced reviews were perceived to be more useful than positively valenced reviews when the product under consideration could be classified as an experience product (i.e. negativity effect), whereas the reverse was observed when the product could be classified as a search product (i.e. positivity effect). As explained by Ahluwalia (2002), a negativity effect can attenuate or reverse into a positivity effect when a product is familiar and liked. In such circumstances, consumers are inclined to defend their liking of a product (even if weak), by giving more scrutiny to positive information about the product. The positivity effect is likely to become prevalent in the situation where reviews discuss search products for which performance evaluations can be made prior to the purchase because of consumers’ familiarity with the products’ attributes. Although the positivity effect was not expected, it provides stronger support for the suggestion that positive and negative review content instigate different effects for search versus experience products because of differences in pre-purchase performance veracity (Park & Kim, 2008; Xia & Bechwati, 2008).

Beyond review valence, argumentation also appeared to be an important predictor of the perceived usefulness of reviews. Reviews that are marked with high levels of argument density and diversity are perceived as more useful. This finding extends previous studies that focused on heuristic cues to explain the effects of reviews, that is, on characteristics shown at the surface level of reviews (e.g. star rating, reviewer identity disclosure) that can be processed with minimal cognitive effort (Chevelier & Mayzlin, 2006; Ghose & Ipeirotis, 2008). The finding that consumers use argument density and diversity to gauge the usefulness of a review provides initial support for the notion that consumers pay attention to characteristics that are more central to the content of the review and that require more elaborate processing (Petty & Cacioppo, 1984).

Finally, we found a positive, albeit relatively weak relation between expertise claims and the perceived usefulness of reviews. This seems to counter the work of Tan and colleagues (2008) on the effects of source expertise on people’s perceptions of online political discussants. The results revealed no
significant relation between expertise and perceived message informativeness. A possible explanation for this differential effect of expertise may lie in its conceptualization: while the study of Tan et al. used status cues to represent a source’s level of expertise (i.e. cues provided by a website to indicate a person’s experience with the website), the present study used expertise claims (i.e. claims provided by the reviewer to indicate his/her experience with the subject), which involves more relevant expertise to the object of discussion. As asserted by Biswas, Biswas, and Das (2006, p. 19) “expertise is topic-specific”; a source must possess knowledge on a particular topic rather than a generalized level to be perceived as an expert.

Implications of the Findings
By demonstrating that review message characteristics make additional contributions to the perceived usefulness of reviews above and beyond the general characteristics shown at the surface of reviews, our research offers several implications. First and foremost, this study makes a theoretical contribution to the literature by showing that the effects of online reviews might not be as straightforward as suggested in the literature. Consumers attach different weights to different reviews depending on which content characteristics are present and which products are evaluated. This means that the use and effects of online consumer reviews cannot and should not be generalized.

A methodological implication of this finding is that when tracking online consumer reviews, one should refrain from using recommendation scores in the form of star ratings as a proxy for review valence. While star ratings provide an important contribution, they explain only part of the variance in perceived usefulness, presumably because of the mainly positive values allotted to star ratings. It was observed in this study, as well as in others (Mulpuru, 2007; Resnick & Zeckhauser, 2002), that the vast majority of reviews tend to receive positive recommendation scores. As such, recommendation scores do not offer a lot of information to prospect purchasers.

Moreover, star ratings “fail to convey important subtleties of online interactions” (Resnick, Zeckhauser, Friedman, & Kuwabara, 2000, p. 47). One such important subtlety involves the proportion of positive versus negative reasoned statements in the open-ended comments of reviews (i.e. argument diversity). The present study shows that variation in the valence of reviews is as important as the overall valence of a review in predicting the review’s perceived usefulness. Such dynamics can be revealed only by analyzing the textual content of a review rather than its star rating.

A practical implication of the finding that review effects cannot be generalized concerns the importance for review sites to develop effective mechanisms that help consumers to gauge information reliability and that enhance consumer trust. Our study provides an empirically based set of tools that may help to unleash the full potential and benefits of information sharing on consumer review sites. For example, based on the finding that argument diversity in reviews is positively related to the perceived usefulness of those reviews, website developers might want to adopt a review format in which reviewers are asked to voice their opinions in a structured way that considers both the positive and negative points of a product.

Limitations and Future Research
The present study responded to the various calls to use naturally occurring consumer interactions on the internet (e.g. Schindler & Bickart, 2005). Although this provides a rich insight into the differences in the content of consumer reviews and the consequences of these differences for the perceived usefulness of reviews, our approach has several limitations, most of which are due to the nature of the data used. One such limitation is that the sample of reviews analyzed for this study were derived from one particular online review site: Amazon.com. Future research should examine whether similar findings will emerge in other online review sites. This is particularly important since message evaluations can be simultaneously affected by a chain of sources. Recent research suggests that people may evaluate online
messages in reference not only to the individual contributor of that information, but also to the website the message derives from (Bronner & De Hoog, 2011; Hu & Sundar, 2010).

Second, the naturalistic design of the present study necessitates some caution in making causal inferences. To control for spurious relationships, we made a good effort to isolate the effects of content characteristics from those emanating from third variables. Despite these efforts it is possible that other, unmeasured variables have affected the results. For example, the design of this study did not allow us to measure variables related to the consumers who have casted useful scores, like consumers’ involvement with the reviewed product. Such variables may be important to take into account, since message effects are generally agreed to result from an interaction between source characteristics, content characteristics, and receiver characteristics (MacInnis, Moorman, & Jaworski, 1991).

A final consideration for future research is the relationship between perceived usefulness and consumer behavior. This study used perceived usefulness as a reflection of review diagnosticity, i.e., the degree to which a review is considered to be useful in the consumer purchase decision-making process (cf. Mudambi & Schuff, 2010). This measure does not capture purchase decision-making per se. Although prior research has found a positive relation between perceived review usefulness and purchase intention (Cheung et al., 2008), additional research is needed to test whether the conclusions of this study can be extended to purchase behavior.

The lessons learned here are important despite the questions that remain. This study suggests that people use several aspects of an online product review in judging the merit of its recommendation, and provide empirical evidence to document the content characteristics that make online reviews a useful source of consumer information. As this is an endeavor that had not previously been accomplished, the present study serves as a springboard for the development of future research directions.

Acknowledgment

This research was partially supported by a Toptalent grant of the Dutch Science Foundation (NWO). The authors wish to thank Nel Ruigrok en Wouter van Atteveldt of LJS Media Research for supporting the data collection.

Notes

1 To cast a vote, one is required to have a password-protected Amazon.com account which is used for at least one purchase to verify account holder’s identity through credit card information. By asking consumers to log into the account before casting a vote, Amazon prohibits users to vote more than once on a review or to vote on their own review.

2 The values $-0.5$ and $0.5$ were applied in situations where conditions were expressed (i.e., if A, then B), nuanced statements were made (identified by such terms as ‘possibly’, ‘maybe’, etc, or future happenings were described.

3 Traditional metrics for intercoder reliability (e.g. Krippendorff’s alpha and Cohen’s kappa), assume that each coder assigns one of a number of possible categories to each unit of observation, and calculate the degree of agreement corrected for chance based on marginal frequencies. Such metrics are difficult to apply to network analysis as each possible relation serves as a unit of observation and the assigned association ($-1 \ldots +1$) as the observation. This results in a large set of units of observation, of which many have a missing value as category. In such cases, the marginal frequencies used for correcting for chance agreement are not proper indicators of relations, leading to widely divergent scores for similar networks. To overcome this problem, we
used a measure from computational linguistics called the F1 score which can be directly compared to alpha or kappa values. To obtain the F1-score, a computer extracts a multiset of extracted relations—i.e., a network—from a given text. Reliability is then assessed by comparing the automatically generated network created by a human coder. This logic was extended to compare networks as generated by two or more human coders (see Van Atteveldt, 2008).

4 With real name, we refer to a registration procedure that Amazon provides for users to indicate their actual name by providing verification with a credit card.

5 The data collected were examined for violation of outlier contamination before performing the OLS regression analysis. Using Cook’s D (critical value: 4/N), 11 cases were identified as outlying cases and, hence, omitted from further analysis.

References


**About the Authors**

**Lotte M. Willemsen** is a PhD Candidate in the Amsterdam School of Communication Research (ASCoR) at the University of Amsterdam. Her research interests include new media, electronic word of mouth, online persuasion, and consumer behavior.

**Fred Bronner** is full professor Media and Advertising Research in the Amsterdam School of Communication Research (ASCoR) at the University of Amsterdam. He also serves as an advisor of TNS NIPO. His main interests are multimedia synergy, consumer decision making, electronic word of mouth and consumers’ economizing tactics.

**Peter Neijens** is full professor of Persuasive Communication in the Amsterdam School of Communication Research (ASCoR) at the University of Amsterdam. His research interests include media & advertising, and persuasion & entertainment.

**Jan A. de Ridder** is president director of the Amsterdam Court of Audit. Before, he was an associate professor Organizational Communication in the Amsterdam School of Communication Research (ASCoR) at the Universiteit van Amsterdam and chair of the department of Communication at that university. His main scientific interests include methods of social science, especially content analysis, organizational communication and political communication. Currently, his research is focused on the performance of governmental institutes.

All correspondence can be addressed to Lotte M. Willemsen, Amsterdam School of Communication Research (ASCoR), University of Amsterdam, Kloveniersburgwal 48, 1012 CX Amsterdam, The Netherlands. Tel: +31 20 525 3191, Email: l.m.willemsen@uva.nl.