International Journal of Business and Management; Vol. 10, No. 7; 2015
ISSN 1833-3850 E-ISSN 1833-8119
Published by Canadian Center of Science and Education

An Empirical Investigation of Self-Selection Bias and Factors Influencing Review Helpfulness

Einar Bjering[1], Lars Jaakko Havro[1] & Øystein Moen[1]

[1] Department of Industrial Economics and Technology Management, Norwegian University of Science and Technology, Norway

Correspondence: Øystein Moen, Department of Industrial Economics and Technology Management, Norwegian University of Science and Technology, Alfred Getz vei 3, 7491 Trondheim, Norway. E-mail: Oeystein.moen@iot.ntnu.no

Received: March 24, 2015 Accepted: May 5, 2015 Online Published: June 20, 2015

doi:10.5539/ijbm.v10n7p16 URL: http://dx.doi.org/10.5539/ijbm.v10n7p16

Abstract

This paper build on 1 489194 product reviews from 30 product categories retrieved from Amazon.com. Product categories are classified by use of natural language analysis tools with computing of subjectivity scores reflecting a search/experience product dimension. Results show a distinct effect of self-selection where the average review score gradually decreases. For most products, no undershooting period was observed, even though a limited number of products groups had this development pattern. Review length, verified purchase and use of real names contributed to increasing helpfulness ratings. The results further suggest search products to be more influenced by review length than experience products.

Keywords: online product reviews, self-selection bias, undershooting, review helpfulness, search and experience products, natural language processing

1. Introduction

Throughout history, people have communicated with each other about experiences with a product, service or a seller (Dellarocas, 2003). *Word of mouth (WOM),* the phenomenon of unsolicited advice between individuals, is widely acknowledged as an important aspect of a product's success (Sundaram et al., 1998). Not surprisingly, scholars and practitioners have long argued that WOM is one of the most effective marketing tools available (Arndt, 1967; Trusov et al., 2009).

Traditionally, consumers have exchanged word-of-mouth through face-to-face conversations. However, with the advent of the internet, this is changing. The internet has brought with it a plethora of new channels from which to discuss the quality of or experience with a product or service (Dellarocas, 2003). As consumers increasingly use the internet to communicate with other consumers, *electronic word-of-mouth (eWOM)* has gained importance (Trusov et al., 2009; Jimenéz & Mendoza, 2013). One particularly important channel for the spread of consumer opinions, is websites providing a platform for consumers to write their own product reviews (Hennig-Thurau et al., 2004). The review system can be a pure review portal, unaffiliated with any merchant or service-provider such as Yelp or TripAdvisor, or it could be included in the functionality of the online merchant or service-providers website directly like Amazon.

US nationwide surveys report that the writing, usage and trust in online consumer generated reviews are growing (Nielsen, 2013; Bazaar Voice, 2013; Pew Internet & American Life Project, 2010) and that consumer opinions posted online are currently the third-most trusted form of advertising (Nielsen, 2013). Therefore, it will become increasingly important for both consumers and businesses to understand how these online product reviews affect the marketplace. Consumers can theoretically be better informed, resulting in more efficient markets (Dellarocas et al., 2006). However, they are also faced with an increasingly complex world of opinions and possibly new challenges, such as how to determine the credibility of a review. If the reviews are skewed or manipulated in any way, consumers could be left confused and it would lead to suboptimal choices (Hu et al., 2012).

In this study we will present different hypotheses divided in two parts. The first group focus how different bias effects influence online product reviews while the second groups of hypotheses highlight how different factors may increase the helpfulness of reviews. Prior research has often used a limited number of reviews, limited

number of product categories or limited number of products. This has reduced the opportunity to develop robust conclusions. We have been able to include 1 489194 reviews (1 147 488 unique) and 4 600 products in 30 different product categories in the study.

2. How WOM Works: The Consumer Socialization Framework

In order to understand how WOM communication influences a consumer's perception of a product or service, and how this ultimately can play into a purchase decision, we will present the *Consumer Socialization Framework* (Figure 1) as described by Wang et al. (2012). The framework consists of three main parts–the antecedents, the process and the outcomes. The aim is to explain, from beginning to end, how consumer socialization affects an individual's attitude towards a product, and eventually, the purchase intention. A central aspect of the model is the type of *social agent* a consumer will encounter in a socialization process. Peers are according to the authors the primary socialization agents, and interaction with these about consumption patterns will greatly influence attitudes towards products and services.

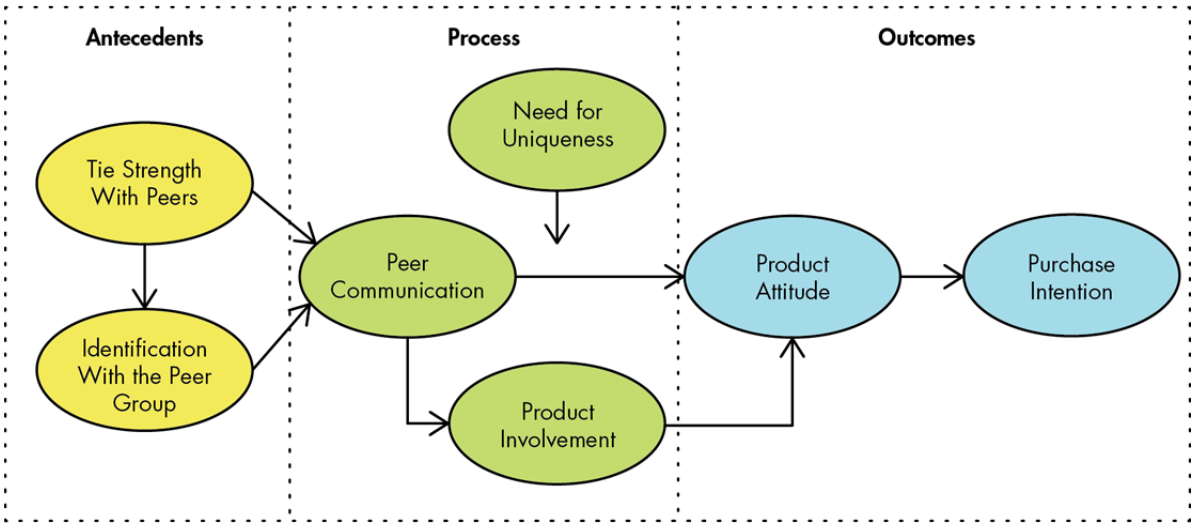

Figure 1. The consumer socialization framework (Wang et al., 2012)

As shown by several other authors, strong ties are perceived as more influential on the individual and produce more cohesion in groups than weak ties (Brown & Reingen, 1987; Granovetter, 1973; Dichter, 1966). Wang et al. (2012) contend the same, positing that one of the antecedents for consumer socialization is the *tie strength with peers*.

The second antecedent, which is influenced by the former, is *identification with the peer group*. In other words, the degree to which an individual conforms to values and beliefs in a group, and identifies with other members of it.

The antecedents both form and influence the first part of the process: *peer communication.* The communication between consumers shape values, attitudes and skills. A moderating effect on the peer communication is the notion of a *need for uniqueness.* Not all individuals respond equally to the normative effects of group communication. The authors separate between low-uniqueness and high-uniqueness consumers; the former will likely be more influenced by peer communication than the latter. Moreover, high-uniqueness consumers may even present "counter conformity motivations" if they feel their identity is threatened.

Product involvement refers to the informative influence from groups (as opposed to the normative), which consists of information and learning a consumer derives from group interaction. Simply by observing or hearing about the use of a product, the consumer may infer a positive attitude or curiosity towards a product or service.

The product involvement and peer communication (moderated by the need for uniqueness) serve to shape the *product attitude*-the way a consumer perceives a product. Consumers having been socially conditioned to prefer a product, or having had higher levels of product involvement can present a more positive product attitude (Wang et al., 2012). Conversely, a group's dislike of a product can influence the consumer to show a negative product attitude.

The product attitude is ultimately believed to be positively associated with *purchase intention*; a positive product attitude will increase the probability of the consumer conducting a purchase. The framework culminates in consumers making a decision of purchase, influenced by the underlying WOM communication with peers.

In the next sections we will development hypothesis were the consumer socialization framework will be a frame of reference.

2.1 Hypothesis Development Part 1: The Bias Focus

When investigating the effects of online reviews and connected communities, it is important to understand what motivates people to read and write reviews. This may have implications for the types of messages that are posted, as well as how they are perceived. In addition, a potential mismatch between reviewers and readers could offer fundamental understanding of the dynamics of online consumer reviews.

Several researchers have looked into the validity of the ratings as a viable indicator of a product's quality (Hu et al., 2006; Koh et al., 2010; Li & Hitt, 2008). Findings indicate that there may be significant limitations to online reviews, due to biases in both reviewers as well as readers. This may be caused by a) Demographic mismatch, b) Consumption goal bias, c) Under-reporting bias and d) Self-selection bias. If we consider the consumer socialization framework presented, review bias in different forms may influence antecedents, processes and outcomes of the interaction.

One source of bias is that of demographic mismatch between reviewer and reader groups. Using the survey data from Pew Internet & American Life Project (2010) Punj (2012) takes a deeper look at the data to try and identify consumer characteristics for specific behavioral groups. He finds that there is a strong demographical mismatch between the group of consumers that posts online reviews, but does not read them, and the group that reads online reviews, but does not post them. The typical exclusive poster is found to be individuals with low education and low income. The exclusive reader group is instead filled by highly educated, well-to-do consumers. Looking at age and gender, the exclusive posting behavior is most commonly found in older consumers, whereas the exclusive reading behavior is more common among young consumers. Although the groups with these behaviors do not define the entire population, the stark differences between them can imply that in certain cases the flow of information will go from consumers that have not conducted research to those that have. This mismatch can reduce the accuracy of an online review, as criteria for assessing products and services may be different between reviewer and reader. Punj (2012) focused exclusive posters and readers while Dellarocas et al. (2007) used data from Yahoo! Movies and found that 74% of the posted reviews in their sample are written by men, in contrast to the fact that only 49% of US moviegoers are male. In addition, the dominating age group is those between 18 and 29, making up 58% of the posted reviews. Again, this does not match the distribution of moviegoers, listing 35% in that age group. This can likely be ascribed to the population using movie portals online. Indeed, Dellarocas and Narayan (2006) see that movie genre is a strong indicator on the *review density,* i.e. the amount of reviews posted per ticket sold. Science fiction movies have a significant positive effect in moviegoers' propensity to post online ratings, whereas comedies and children's movies have a significant negative effect. Both indicate younger males being more active in reviewing movies online, and suggesting that product category has a strong influence on the reviewer demographics.

Biases may also exist in consumers doing product research. Zhang et al. (2010) propose that the *consumption goals* that consumers associate with the reviewed product moderate the effect of review valence on persuasiveness. Consumption goals are categorized as either promotion goals or prevention goals. Promotion goals are fit for scenarios in which consumers wish to use the product to reach some positive end-state, e.g. using photo-editing software to improve a photo. Prevention goals are common for products that are used to avoid a negative end-state, e.g. antivirus software. The study finds that consumers who evaluate products associated with promotion consumption goals perceive positive reviews to be more persuasive than negative ones (i.e., a positivity bias). Conversely, consumers evaluating products associated with prevention consumption goals perceive negative reviews to be more persuasive than positive ones (i.e., a negativity bias). This implies that even though the average rating may be objectively correct, consumers may infer a different score depending on the consumption goals.

The third and most commonly cited shortcoming of online reviews is *under-reporting bias*. This is likely primarily a consequence of the motivations for posting reviews, in which extremely satisfied or extremely dissatisfied consumers are more likely to post reviews. Consumers with mediocre or average experiences simply don't find the same utility in expressing their views (Anderson, 1998). As such, the rating distributions approach the U-shaped curves, where the average values are underrepresented. In fact, Hu et al. (2006) find that about 53% of reviews posted on Amazon.com are bimodal, showing signs of the U-shape. As these product ratings do not

have a normal distribution, they do not necessarily have credible mean values, and may be misleading for consumers.

Self-selection bias is a phenomenon that occurs when products have a subset of consumers that are especially invested in that product, its producer or category (Li & Hitt, 2008). For instance, if an author has a loyal following, it is likely that this subset of consumers will be strongly represented in the early adopters of new books. The first reviews may therefore be positively biased, misrepresenting the true quality of the product until a sufficiently large amount of unbiased reviewers pitch in. Li and Hitt (2008) find that about 70% of the reviewed books on Amazon show signs of self-selection. Further evidence of the bias has been reported in several instances (Dellarocas et al., 2007; Hu et al., 2011; Zhu & Zhang, 2010). Then, our first hypothesis is:

H1: The average rating of a product tends to decrease over time before stabilizing at a long term value lower than the initial value.

The self-selection of early reviewers may also cause secondary problems. When disappointed consumers, having bought into the biased early reviews, post their experiences, they over-compensate and post reviews that are more negative than the average long term value (Li & Hitt, 2008). Known as the *undershooting period*, this is found in about 20% of the books reviewed on Amazon. Li and Hitt (2008) find that products that are affected by undershooting on average see the dip in ratings between the 6th and 19th weeks after release. This leads us to the next hypothesis:

H2: Some products with a difference in initial average rating and long term average rating go through an "undershooting" period after the initial period where the rating is lower than the long term average.

Li and Hitt (2008) show that the undershooting effect is stronger for products with more heterogenous consumer preferences. As previously argued, experience products are believed to be evaluated more on subjective preferences than search products. For products being evaluated purely on objective criteria, it is easy to imagine consumer preferences to converge; either the product matches the objectively defined need or usage, or it doesn't. However, products being evaluated on a basis of subjective taste will likely see more diverging preferences. Li and Hitt (2008) found support for their hypothesis based solely on data from books, looking at differences in consumer preferences for products within that category. However, considering the assumed dynamic between search and experience products, it seems plausible that differences in the magnitude of any undershooting can also be observed between product types. Specifically:

H3: The undershooting effect is stronger for experience products than for search products.

2.2 Hypothesis Development Part 2: The Ties Importance

In the consumer socialization framework we notice that tie strengths and identification with the peer group are important antecedents. It is possible that characteristics of the reviews will influence the perceived ties and impact of the review.

Specifically, we will look at a particular metric provided by several online retailers with reviews, namely that of review helpfulness. As described, consumers are often asked if they find a particular review to be helpful or not. The number of helpful votes are then displayed right next to or below the review, giving consumers some added information of the quality of specific reviews. The review helpfulness has been shown to be influential for the impact of reviews on several occations (Hu et al., 2008; Forman et al., 2008)

A previous analysis of reviews from Amazon.com across six products indicated that review extremity, review depth, and product type affect the perceived helpfulness of the review (Mudambi & Schuff, 2010). Independently of product type, review depth, or rather, the length of the review measured in word count, was found to have a positive effect on the helpfulness of the review. This is simply believed to stem from the increased information provided in the review. Longer reviews are assumed to include more nuanced descriptions of the products, considering both positive as well as negative sides (Chevalier & Mazlyn, 2006). However, Mudambi and Schuff (2010) found that the effect was greater on the helpfulness of the review for search goods than for experience goods. Since reviews for search goods are often presented in a fact-based, reviews can be relatively short. The factual nature of search reviews implies that additional content in those reviews is more likely to contain important information about how the product is used and how it compares to alternatives. While additional review content is helpful for all reviews, Mudambi and Schuff (2010) found that the incremental value of additional content in a search review was more likely to be helpful to the purchase decision than the incremental value of additional content for experience reviews.

Expanding on previous findings, we expect to see similar effects:

H4a: Reviews that are perceived as helpful tend to be longer than other reviews.

H4b: The association between helpfulness and the length of the review is stronger for search goods than for experience goods.

One trend that has been seen amongst online retailers, is the growing inclusion of social network functionality. This development finds support in research. Forman et al. (2008) found that online community members rate reviews containing identity-descriptive information more positively, and that the prevalence of reviewer disclosure of identity information is associated with increases in subsequent online product sales. Reviewers who disclosed real name or location had 12.2 percentage points more helpful votes than otherwise identical reviewers. Wang (2010) found that there were an order of magnitude more reviewers on Yelp than for two competing sites, prolific referring to the productivity and perceived helpfulness. This is believed to be a consequence of increased trust in the reviewers stemming from Yelp's encouragement of creating social profiles on their review system. Wang (2010) contends that this trust is critical for consumers when assessing reviews. We formulate our next hypothesis.

H5: Reviews written by reviewers that use their real name are perceived as more helpful than other reviews.

Considering the consumer socialization framework, this hypothesis suggest that using real names increase the tie strenght (trust) between reviewers and readers. Another element that may increase trust in the reviews is a verification that a transaction has taken place. If consumers can be certain that the reviewer actually has used the product, a greater amount of credibility can be attributed to the opinions in the review. Many online retailers display such information (e.g. Amazon.com), whereas some require a purchase for the consumer to be able to review the item (e.g. Hotels.com). Consumers are increasingly wary of review manipulation and fake or shill reviews (Bambauer-Sachse & Mangold, 2013), so it should be conceivable that reviews with verified purchases should see larger amounts of trust, and thus be perceived as more helpful. We postulate:

H6: Reviews written by reviewers with verified purchases are perceived as more helpful than other reviews.

3. Methodology

We build on models presented in prior research as Li and Hitt (2008) and Mudambi and Schuff (2010). The data source is Amazon.com, they are the world's largest online retailer and includes millions of products in hundreds of categories. We used the API (Application Programming Interface) offered by Amazon and developed JAVA scripts in order to perform the data retrieval from the Amazon system.

30 different product categories were selected, including the 100 best selling products in each of these categories (as of March 19, 2014). In addition, 100 random products (ex the 100 best selling products) were randomly picked in each category. The random products were included, as only focus on the bestselling products would reduce the opportunity to generalize the results independent of product popularity. In addition, the most popular products were expected to be more exposed to other aspects as marketing campaigns and media attention. In practice, the random products serves as a control group, they were chosen by a Java script based on random words (Wordnik/get Random Words).

3.1 Collecting Reviews and Description of the Dataset

We used a VBA script to download the necessary data from Amazon. In total, we retrieved 1 489 194 reviews. Of these, 1 147 488 are primary reviews from 30 product categories, table 1 present key information about the numbers and categories. The distinction between primary and secondary reviews is caused by duplicates where primary reviews are unique. Secondary reviews are copies or duplicates of one in the primary, but have a non-duplicate ASIN (Amazon Standard Identification Number). This is due to the overlap between versions of products, in examples a movie may be relased in different platforms as DVD or Bluray where each version gets a unique ASIN but in fact refers to the same basic product with shared reviews.

Considering the primary reviews, they include 986 344 reviews of the top 100 products and the 100 random products have 161 144 reviews. Variation between categories is large, with variation from under 2 000 reviews (hobby fabrics 1 838; screws 1 919) to 100 000+ (movies 103 586) and more than 200 000 (books 216 361). Evaluating the mean number of reviews movies has the highest and envelopes the smallest. 99% of the reviews have less than 2 705 characters, the average length is relatively short (160 characters). The length mode is even shorter (between 111 and 121) for all product categories indicating the existence of a limited number of long reviews. As the mean length varies from 650 (digital cameras) to 205 (jewelry) we notice the occurance of these long reviews varies between product categories. In addition, the rating of the products are uneven as mean value is 4.27 and 64.9% of ratings have five stars. Then, it is a distinct pattern that many reviews are positive.

Table 1. Number of primary and secondary reviews for all categories

Category	Primary	Secondary	All
Board Games	34 407	395	34 802
Books	216 361	0	216 361
Bowls	4 116	2 832	6 948
Can Openers	22 068	4 984	27 052
Candy	13 793	1 973	15 766
Car Electronics	93 027	19 822	112 849
Clothing	68 466	0	68 466
Copy Paper	3 352	743	4 095
Desktop Computers	5 594	390	5 984
Digital Cameras	41 132	8 601	49 733
Dog Food	15 697	3 071	18 768
Envelopes	3 189	97	3 286
Guitars	9 239	5 293	14 532
Hardware	40 094	293	40 387
Hard Drives*	19 375	24 304	43 679
Hobby Fabric	1 838	265	2 103
Ink and Toner*	27 581	4 671	32 252
Jewelry	21 530	0	21 530
Ladders	12 960	3 390	16 350
Movies	103 586	127 393	230 979
Perfumes	21 724	358	22 082
Restroom Fictures	2 735	968	3 703
Screws	1 919	120	2 039
Shoes	59 652	0	59 652
Software	41 169	10 400	51 569
Test and Measure	18 186	110	18 296
USB Drives*	48 272	97 432	145 704
Video Games	90 018	19 724	109 742
Vitamins	73 816	5 017	78 833
Watches	31 652	0	31 652
Total	**1 147 488**	**341 706**	**1 489 194**

Note. *No random product list generated.

3.2 The Subjective Score

The distinction between different types of products is included in some of the hypotheses, dividing between search and experience products. Just using two categories will underestimate the differences between products as this may be regarded as a scale where products are located at different points. In order to classify products, we used the subjectivity of reviews as an indicator expecting search products as USB drives or copy paper to have less subjective sentences in their evaluations compared to experience products as books or movies.

As our method for measuring subjectivity we used computer based sentiment analysis employing the OpionFinder library developed at the University of Pittsburgh, Cornell University and the University of Utah. In prior studies, the OpionFinder classifiers have been used with good results in terms of classifying subjectivity (He et al., 2008). The toolkit used (subjectivity classifiers) are based on work of Riloff and Wiebe (2003) and Wiebe and Riloff (2005).

The OpinionFinder toolkit has two different subjectivity classifiers. One is model based and trained by machine learning. It has a reported accuracy of 76% and classifies all sentences as subjective or objective. The other model is rule-based, applying defined rules to classify whether a sentence is subjective or objective. It is reported as having a higher accuracy (91.7% for subjective sentences, 83% for objective sentences) but it only classifies sentences if it is able to do this with confidence leaving sentences also as unclassified. In the calculation of subjective with the rule based approach we excluded unclassified sentences and calculated the number of subjective sentences divided with the number classified as either subjective or objective.

The calculated subjective score rate produce results as expected with highest subjectivity score for books and

movies and lowest values for hard drives and copypaper.

3.3 Model Specification

In the hypotheses, we look at the expected self-selection bias as well as over- and undershooting periods. In order to assess the existence of self-selection bias and review undershooting, Li and Hitt (2008) developed an enhanced negative exponential model to fit the trend in book reviews over time. Their model was formulated as follows:

$$AVGRATINGpt = f_0 + f_1 \exp(-f_2 * T_{pt}) \cos(f_3 * T_{pt}) + u_p + e_{pt} \quad (1)$$

AVGRATING represents the average rating of all reviews posted for product p between the time it was released and time t. T denotes the amount of three-day intervals that have passed since release at time t. u_p represents a fixed effect and e_{pt} denotes a random error. Both these variables are included for symbolic representation, because as we are not able to observe or measure them, they will not be included in the actual calculations.

The coefficients, f_1, f_2 and f_3 illustrate the trend in the average rating over time. Depending on the signs of f_1 and f2, the model can display an increasing ($f_1f_2<0$), decreasing ($f_1f_2>0$) or no trend ($f_1f_2 = 0$) over time (Li & Hitt, 2008). The test for self-selection is thus simply an assessment over the signs of f_1 and f_2. A positive sign will imply the existence of self-selection.

The model also includes a cosine term. If the final coefficient (f_3), is zero, we see that the cosine term equals 1, and the model becomes a standard negative exponential model (Li & Hitt, 2008). However, if f_3 is non-zero, the cosine term will produce the revealing "dip" characteristic to undershooting.

For the hypotheses 4, 5 and 6, the dependent variable becomes helpfulness of the review. Previous literature (Mudambi & Schuff, 2010; Forman et al., 2008) suggests that helpfulness is moderated by review depth, product type as well as disclosure of reviewer identity and purchase verification. We formulate a regression model to incorporate these variables:

$$PCTHELPFUL = \alpha + \beta_1 RATING + \beta_2 CHARCOUNT + \beta_3 PRODUCTTYPE + \beta_4 CHARCOUNT\,x\,PRODUCTTYPE + \beta_5 IDENTITY + \beta_6 VERIFIED + \varepsilon \quad (2)$$

PCTHELPFUL denotes the percentage of votes awarded to the review deeming it helpful. RATING refers to the rating awarded by the speficic review, and CHARCOUNT represents the total number of characters typed in for the review. PRODUCTTYPE is a binary (dummy) variable coded to 1 or 0, with 1 indicating an experience product and 0 indicating a search product. The CHARCOUNT x PRODUCTTYPE aims to catch the effect sought that the review depth is more important for helpfulness of reviews for search products than for experience products. IDENTITY is a binary variable indicating whether or not the reviewer has chosen to disclose his or her identity. The final variable, the dummy VERIFIED, indicates whether or not the review is connected to a verfied purchase.

This model use traditional multiple linear regression as the method of analysis, where the basis from Mudambi and Schuff (2010) used a Tobit regression. This hinged on the notion that the review helpfulness is bounded in its extremes, and the sample is censored in nature. Consumers can only vote helpful or not helpful, and there may be self-selection issues in who actually decides to vote. However, for robustness, they also test their model as a multiple linear regression. This gave the same findings and significances.

4. Results

To test for occurrences of self-selection in our database of reviews, we first run the enhanced negative exponential model on the whole set, as Li and Hitt (2008). Products with less than three reviews were removed, in order to require each product to have a meaningful average. This limits the total amount of products to 4,342, which renders the following optimal solution:

$$AVGRATING_{pt} = 4.102 + 0.290 \exp(-0.04 * T_{pt}) \cos(7.126E-6 * T_{pt}) + u_p + e_{pt} \quad (3)$$

The equation shows a clear downward trend for the set as a whole, as seen in Figure 1. This supports the notion that self-selection is a phenomenon that occurs with online consumer reviews. The cosine term is quite small, which means the characteristic self-selection dip is not immediately visible when regressing over all products at once. This mirrors the findings in Li and Hitt (2008). A more detailed analysis is required to assess the prevalence of undershooting, however.

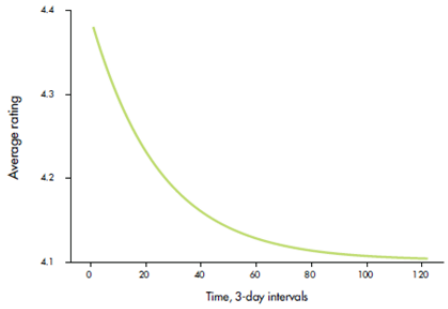

Figure 1. Average rating plotted against time

When repeating the regression for each individual product, a significant amount did not have enough review activity in the first several time intervals, resulting in failed regressions. 866 products thus returned no results, leaving us 3,476 completed regressions. The results of the product-level regressions are listed in Table 2.

Table 2. Regressions, self-selection and undershooting at the product level

	Declining	**Increasing**	**Undershoot**	**Not sign.**
Number	293	104	240	3.079
Percent	8.43	2.99	6.90	88.57

Note. N=3746, significance measured as $p<0.1$.

As we can determine from the table, the overwhelming majority, or nearly 90%, of regressions did not return any significant results. This is likely a consequence of fixed effects and other noise in the data, as well as inaccurate starting values for the regression parameters.

We do find several individual instances of declining curves for average rating, which supports hypothesis 1, predicting the existence of self-selection. Out of the 397 cases with significant trends (either increasing or declining rating averages), 293 products, or roughly 74%, display evidence of the self-selection effect. 104 cases, or approximately 26% show a growing trend.

Out of the 293 products that show a declining trend, 240, or almost 82% also have an f_3 parameter that differs significantly from 0, meaning they test positive for undershooting. How many of these that actually show visible undershooting is not entirely clear, however, since many hold relatively small values. In addition, comparing with the results reported by Li and Hitt (2008) is difficult, since they do not disclose any threshold value. Notwithstanding, the existence of undershooting finds limited evidence, indicating no such effect for most products and we reject this hypothesis (H2).

We find mixed support for our hypotheses regarding review distributions and biases. We find evidence of self-selection bias in the dataset as a whole, but product-specific regressions only return a few results. We find some evidence of the existence of undershooting, but are not able to make estimations of its prevalence. We also have insufficient results to infer any differences across categories.

For testing of the hypotheses 5 and 6 we build on the work by Mudambi and Schuff (2010). When structuring the data for these tests, it soon became obvious that there was a strong bias towards reviews with 100% or 0% helpful votes. These mostly consisted of cases with single votes of either helpful or not, severely skewing the results without having much data to back it. Therefore, we pruned these cases from our data, which resulted in a much more even distribution, resembling a gaussian curve. This still includes all recorded values between a perfect 0 and 100%, which should suffice to determine any effects for review helpfulness in general.

For our first regression, we simply replicate the test as performed by Mudambi and Schuff (2010), albeit with a few omitted interaction terms describing the effects of rating with product category as well as the square of the rating. As these interaction terms did not affect neither the sign nor magnitude of our focus variables, they were omitted in order to reduce contextually redundant information. Results of the regressions are shown in table 3 and tests the suppositions described for hypotheses 4, 5 and 6.

Table 3. Regression results: helpfulness and product category

Variable	Coeff.	Std.Err.	Std.Coeff.	T	Sig.
(Constant)	48.011	0.294	-	163.361	.000
Rating	2.792	0.035	0.220	79.649	.000
Total votes	0.024	0.001	0.065	23.295	.000
Ver. Purchase	2.308	0.119	0.055	19.34	.000
Real name	1.304	0.137	0.026	9.52	.000
Chars.	0.005	0	0.246	16.744	.000
Type	-8.694	0.271	-0.113	-32.085	.000
Type x chars	-0.002	0	-0.082	-5.481	.000

Note. DEP: Helpfulness (%), $R^2 = 0.098$, N = 120,552.

We see several expected effects. The two most prevalent of which are the effect of the review's star rating, as well as the effect of the length of the review, measured in the amount of characters. The signs are both positive, indicating that a more positive rating, as well as a longer review, tends to see higher percentages of helpfulness.

We also note that the coefficients for the verified purchase and real name badges are positive and significant. Consumers seem to appreciate information that lessens the risk of fraudulent or manipulated reviews. This is in accordance with the hypothesized effect, and supports hypotheses 5 and 6.

The main focus of the test, however, is the interaction term between product type and the length of reviews, as measured in the number of characters. The sign is negative, supporting hypothesis 4b, which means consumers rate longer reviews more helpful for search products. This effect is attributed to the fact that search product reviews tend to be short and factual in nature. In the discussion section, we will further comment on the relationsship between product type and review length.

Substituting the binary product type variable for our continuous subjectivity variables, we would expect to see similar results as for the first test. Indeed, as shown in table 4 (rule based subjectivity) and table 5 (mode based subjectivity) both variants of the subjectivity classification seem to hold up, lending additional support for hypothesis 4b. Objectively evaluated products are associated with a stronger effect from the length of reviews on helpfulness.

Table 4. Regression results: helpfulness and rule based subjectivity

Variable	Coeff.	Std.Err.	Std.Coeff.	T	Sig.
(Constant)	50.094	0.207		241.805	.000
Rating	2.619	0.03	0.203	88.313	.000
Total votes	0.009	0	0.049	21.422	.000
Ver. Purchase	1.607	0.103	0.038	15.57	.000
Real name	1.833	0.115	0.037	15.946	.000
Chars	0.004	0	0.204	30.563	.000
Subj. (R)	-44.849	1.004	-0.127	-44.678	.000
Chars x subj. (R)	-0.002	0.001	-0.022	-3.177	.001

Note. DEP: Helpfulness (%), $R^2 = 0.100$, N = 172,663.

Table 5. Regression results: helpfulness and model based subjectivity

Variable	Coeff.	Std.Err.	Std.Coeff.	T	Sig.
(Constant)	63.391	0.546		116.098	.000
Rating	2.646	0.03	0.205	88.878	.000
Total Votes	0.009	0	0.500	21.637	.000
Ver. Purchase	2.167	0.101	0.051	21.408	.000
Real name	1.759	0.115	0.035	15.268	.000
Chars	0.006	0	0.285	13.098	.000
Subj. (M)	-53.021	1.367	-0.110	-38.784	.000
Chars x subj. (M)	-0.005	0.001	-0.101	-4.644	.000

Note. DEP: Helpfulness (%), $R^2 = 0.097$, N = 172,663.

Overall, we find support for all hypotheses regarding review helfulness. The length of the review is in all cases associated with larger values of heplfulness. The verified purchase and real name badges also show positive signs in all regressions, all highly significant. Replicating the test for product type by Mudambi and Schuff (2010), we practically mirror their findings, with the exception of the total votes variable, which switches signs. The most important finding, however, is that our subjectivity variable matches the expected results, showing stronger effect of the length of review for objectively evaluated products. Table 6 summaries the results of the hypothesis testing.

Table 6. Summary of results for hypothesis testing

	Hypothesis	**Result**
H1	The average rating of a product tends to decrease over time before stabilizing at a long term value lower than the initial value.	Supported
H2	Some products with a difference in initial average rating and long term average rating go through an "undershooting" period after the initial period where the rating is lower than the long term average.	Rejected
H3	The undershooting effect is stronger for experience products than for search products.	Rejected
H4a	Reviews that are perceived as helpful tend to be longer than other reviews.	Supported
H4b	The association between helpfulness and the length of the review is stronger for search goods than for experience goods.	Supported
H5	Reviews written by reviewers that use their real name are perceived as more helpful than other reviews.	Supported
H6	Reviews written by reviewers with verified purchases are perceived as more helpful than other reviews.	Supported

5. Discussion

The results presented have distinct results increasing our understanding of online product evaluations based on a dataset including a large set of reviews, product categories and products within the 30 categories selected.

5.1 Average Scores Decrease Over Time

First, for people reading reviews and for companies, it is relevant important to understand that average review score will decrease over time. This is not least important as most attention on reviews is likely to exist in the first time periods after product launch where the review scores are most unreliable or different from the likely long term average. In addition to the self-selection bias, two other reasons for this score decrease may exist. First, when consumers increase their product experience (having owned it for longer time) it is possible that weaknesses (like reliability issues) will occur more often than within the first days of use. Then, a higher share of reviewers may have more experience as the time from product launch increases influencing the review content. Second, a new product may be more competitive than older products. Then, expectations and comparisons with other products may reduce the score gradually as (new) product alternatives improves and this may contribute to the decreasing tendency.

5.2 Positive Reviews Are Most Valued by Consumers?

The effect of a positive review might be attributed to how the users approach reviews. A typical consumer will surf review systems in order to help them in a purchase decision. In many cases, one would assume that the consumer has a certain need and is trying to find a remedy to take care of that need. The goal is to find the right product. In this case, a positive review might reassure a consumer with doubt, whereas a negative review might dispel the consumer from making a choice, and in effect, prolonging the search process. Our interpretation of the positive regression coefficient (rating versus helpfulnes in tables 3, 4 and 5) is that consumers value reviews helping them make good decisions more than reviews helping them avoid bad ones.

5.3 What about Review Length?

The interpretation of the positive coefficient for the amount of characters is less convoluted. As outlined, longer reviews tend to contain more information, and often include both negative and positive sides to a product. This implies that consumers value reviews that describe products and their use in detail, rather than short sentiments that do not provide much information.

We found the expected effect of review helpfulness paired with product categories. Search products are assumed

to see larger effects of review length on review helpfulness, since search product reviews tend to be short and factual in nature (Mudambi & Schuff, 2010). As indicated in the results section, this results need some more consideration. In table 7 we display the means when splitting in search and experience categories, respectively. The experience products see approximately 90 characters more per review than search products (374.86 vs. 283.53 characters).

Table 7. Mode and mean of characters per review for search and experience goods, Search goods in the top half, experience goods in the bottom half

CATEGORY	MODE	MEAN
Bowls	116	484.85
Can Openers	113	311.53
Copy Paper	111	263.79
Envelopes	125	239.38
Hard Drives	112	413.41
Ink and Toner	114	253.21
Screws	120	306.89
USB Drives	113	287.15
Board Games	115	394.30
Books	118	433.02
Candy	123	293.43
Movies	113	381.39
Perfumes	114	293.95
Video Games	112	452.99
Vitamins	113	380.09

However, recalling the distribution of lengths of reviews, the vast majority of reviews are quite short, with a minority of long reviews skewing the mean value upwards. The mode, however, is remarkably similar across categories. Further, looking at the 5% trimmed means, where the top and bottom 5% are cut, the difference is only 38.39 (238.80 vs 277.19). This suggests that the effect measured in our tests concerns a relatively small fraction of the sample, specifically those that lie to the right of the mode. Also, it somewhat challenges the assumption that search product reviews are inherently shorter, since it's mostly a case of the longest reviews being longer for certain of the experience products, such as books or video games.

All our tests, as well as those by Mudambi and Schuff (2010) show the positive effect of review length to be larger for search (objectively evaluated) products, but we contend there may be different ways to interpret this. As the simple length argument, asserting that search product reviews are shorter, seems slightly weakened, our suggestion will be to more closely examine differences in the content of the reviews, to see if there may be other factors at play. Mudambi and Schuff (2010) argue that the incremental value of more depth to a search good review will be larger than for a experience good review, basing this on the fact that an increase in subjective information has less informational value. This may very well be the case, and indeed, we have not seen contradictory evidence. However, we propose that one should also consider that reviews for search goods may increase in subjectivity level as the length of the review increases, simply because the objective facts have all been stated. If this is the case, one would have to consider the fact that the increased effect of length on helpfulness may stem from subjective statements holding a higher informational value in an otherwise objective review.

5.4 Total Votes and Helpfulness

Further, we note that the total votes are positively associated with higher helpfulness. This is somewhat in contrast to previous findings (Mudambi & Schuff, 2010). However, our result has a higher statistical significance ($p < 0.001$) but a smaller relative magnitude. We interpret the sign of total votes to be an effect of increased visibility. As a review garners several helpful votes, it may eventually reach the frontpage for that product, where typically the most helpful reviews are presented. These reviews will likely see a massive amount of exposure compared to reviews with smaller amounts of helpfulness. As time passes, these reviews will then see ever larger amounts of total votes, and since they've already been voted up as the most helpful, they will in all likelihood continue to be voted relatively helpful. This process will, according to our understanding, give the most helpful reviews a larger amount of votes, sometimes by significant amounts. This may therefore skew the effect of the

total votes in favour of reviews found most helpful, resulting in a positive coefficient for our total votes variable.

6. Implications for Further Research

Considering further research, we will in particular suggest four major themes that need to be adressed.

6.1 Classification of Products-The Usability of the Subjectiv Measure

One of the effects of eWOW is the increased ability for consumers to evaluate products prior to usage as they will have access to the experience and views of other consumers. As a consequense, there may be a reduced number of products beeing experience products in a strict sense. With large review bases, experience products as books or movies novels will not be evaluted equally by all, but the ability to get a impression will be improved. For research, the limited usability of a simple classification either as search or experience products represents as challenge. Our subjective score variable allow for classification of all products and categories and we notice that the categories scores makes sense compared to expectations. This use of computerized language processing techniques reduces ambiguity and represents an approach making it possible to replicate research and classify product categories better than what have been observed in prior research. Further studies should further focus on approaches for classifying products as differences seem to exist depending on product type.

6.2 Product Complexity

When looking at mean rating we noted that there are some differences across product categories, and suggest this may stem from an unequal risk of misuse of certain products. Especially software stands out with a particularly low rating. Another possible explanation for this could be that it is in fact the product complexity that drives the risk of misuse, and that this is a separate driver for the need of quality information.

Although not their main focus, Chen and Xie (2008) suggest that reviews will have the most effect for complex, or high-tech products, for which it is difficult to attain the requisite knowledge for correct use. Chen and Xie (2008) further argue that novice users will be unsure if the product matches their preferences or suits their needs, and will benefit from reading reviews written by more expert users.

Considering this, we briefly tested one proposed method for quantifying product complexity using the number of distinct words in a review set adjusted for the number of reviews. This notion was based on presumption that complexity requires a wider vocabulary, and that different reviews will explain the complexity differently-while reviews tend to be more alike in word selection for less complex products. However, this classification did not provide any sensible results. Future research should try different approaches in quantifying complexity, and assess whether this is a driver of its own for review impact.

6.3 The Consumer Socialization Framework

Key elements of this framework are tie strength, group identification and communication. All of these factors may be influenced by bias effects as well as review content were in example length, real names and verified purchases may have a positive effect on the trust and reliability as evaluated by readers. Overall, our results fit with this model in line with expectations. In further studies new approaches should be tested. One example could be how self-description (as gender, age, family status, interests) increases helpfulness of reviews or add to a potential impact on buying decisions. In fact, the factors we have included may just be one step in classifying how content influences impact, far more detailed analyses may be possible even being challenging in terms of data retrieval and handling.

6.4 Self-Selection and Popularity–Coinciding Effects?

Prior research have showed a relation between product rating and sales. In hypothesis 4a we present evidence that the products ratings over time in many cases is lowered from the initial average rating due to the phenomenon of self-selection. The 100 best selling products may be new launches and have small possibilities for increased sales, and are more likely to experience a decrease in sales in the forthcoming weeks independent of how the average rating develops. Considered together with the self-selection effect, which predicts that products who are recently released will drop in average rating, one could thus further argue that a link between sales rank and review scores is perhaps simply a coincidence of two different mechanisms (the self selection effect reducing average review score and sales decline gradually after product launch), and that they have little or nothing to do with each other. These mechanisms need further investigation as it may have serious consequenses on our understanding of the impact of online product reviews if there is this type of coinciding effects.

7. Concluding Remarks and Implications

For managers, online product reviews at independent platforms is not something they have large opportunities to influence. Considering our results, they contribute to the understanding of bias effects and why some reviews are

more helpful than others and how this may influence sales. We have focused one bias effect (self-selection) and identified a pattern development of decreasing average rating and a undershooting effect (time with score belowe average long term average) for a few product groups. In addition, we have showed how increased review length also increase helpfulness in particular for search products and how open reviewer identity and verified purchases increases helpfulness. When companies run their own consumer ability to comment on products, one direct implication is that motivation for longer reviews, open reviewer profiles and possibility to verify purchases will increase the helpfulness as evaluated by readers.

Then, there is a question about how a company should handle independent product review sites. Based on our data, 64.9% of all reviews has a five star rating, suggesting a large proportion of positive reviews. When we know that the average score typically decreases, stimulating early product adapters to write reviews on independent sites seems as a possible manager priority. However, this and other ways of influencing independent products reviews sites may be regarded as manipulation or direct misleading consumers. Therefore, we will point to three some priorities not related to review manipulation: First–our data set imply a U-shaped review distribution: both the most satisfied and most dissatisfied customers are more visible than the average satisfied group. In some sites, the most helpful positive and negative review is highly visible. As a consequense–the handling of dissatisfied customers, support systems and how they are met by the company is increasingly important as a very dissatsified customere may use online reviews as a way to express perceived product and service negative views. While this reflects internal improvent processes, our second suggestion is to consider responses to independent negative reviews: In some cases dissatisfaction is caused by usage errors or inability to handle products correctly. Then companies may in fact advice consumers presenting negative reviews on independent platforms. We have observed how companies as Sony are actively guiding and responding to complaints at Amazon.com, attempting to reduce the effect of negative reviews for product categories as cameras and camcorders with an official company profile. In fact, the development suggest that consumer care increasingly should involve responses to negative reviews on sites as Amazon.com. Third: Product reviews may be helpful for consumers but also the comments and suggestions/problems experienced should be systemized and analysed in order to improve product quality and functions-in this process both objective statements as well as more subjective evaluations should be considered.

References

Anderson, E. W. (1998). Customer satisfaction and word of mouth. *Journal of Service Research*, *1*, 5-17. http://dx.doi.org/10.1177/109467059800100102

Arndt, J. (1967). The role of product-related conversations in the diffusion of a new product. *Journal of Marketing Research*, 291-295.

Bambauer-Sachse, S., & Mangold, S. (2013). Do consumers still believe what is said in online product reviews? A persuasion knowledge approach. *Journal of Retailing and Consumer Services*, *20*, 373-381. http://dx.doi.org/10.1016/j.jretconser.2013.03.004

Bazaar Voice. (2013). *Social commerce statistics. Bazaar Voice*. Retrieved from http://www.bazaarvoice.com/research-and-insight/social-commerce-statistics/#Trust

Brown, J. J., & Reingen, P. H. (1987). Social ties and word-of-mouth referral behavior. *Journal of Consumer Research*, 350-362. http://dx.doi.org/10.1086/209118

Chen, Y., & Xie, J. (2008). Online consumer review: Word-of-mouth as a new element of marketing communication mix. *Management Science, 54*, 477-491. http://dx.doi.org/10.1287/mnsc.1070.0810

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

Dellaracos, C., Zhang, X., & Awad, N. F. (2007). Exploring the value of online product reviews in forecasting sales: The case of motion pictures. *Journal of Interactive Marketing, 21*, 23-45.

Dellarocas, C. (2003). The digitization of word of mouth: Promise and challenges of online feedback mechanisms. *Management Science, 49*, 1407-1424. http://dx.doi.org/10.1287/mnsc.49.10.1407.17308

Dellarocas, C., & Narayan, R. (2006). A Statistical Measure of a Population's Propensity to Engage in Post-Purchase Online Word-of-Mouth. *Statistical Science, 21*, 277-285.

Dellarocas, C., Zhang, X. M., & Awad, N. F. (2006). Strategic manipulation of internet opinion forums: Implications for consumers and firms. *Management Science, 52*(10), 1577-1593.

Dichter, E. (1966). How word-of-mouth advertising works. *Harvard Business Review, 44*.

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*, 291-313. http://dx.doi.org/10.1287/isre.1080.0193

Granovetter, M. S. (1973). The strength of weak ties. *American Journal of Sociology,* 1360-1380.

He, B., MacDonald, C., & Ounis, I. (2008). Ranking opinionated blog post using OpinonFinder. In Proceedings of the 31st annual international ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 727-728.

Henning, E. T., Al-Thurau, T., Gwinner, K. P., Walsh, G., & Gremler, D. D. (2004). Electronic word-of-mouth via consumer-opinion platforms: What motivates consumers to articulate themselves on the Internet? *Journal of Interactive Marketing, 18,* 38-52.

Hu, N., Bose, I., Koh, N. S., & Liu, L. (2012). Manipulation of online reviews: An analysis of ratings, readability, and sentiments. *Decision Support Systems*, *52*, 674-684.

Hu, N., Liu, L., & Sambamurthy, V. (2011). Fraud detection in online consumer reviews. *Decision Support Systems, 50*, 614-626. http://dx.doi.org/10.1016/j.dss.2010.08.012

Hu, N., Liu, L., & Zhang, J. J. (2008). Do online reviews affect product sales? The role of reviewer characteristics and temporal effects. *Information Technology and Management, 9*, 201-214.

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. ACM, 324-330.

Jimenez, F. R., & Mendoza, N. A. (2013). Too Popular to Ignore: The Influence of Online Reviews on Purchase Intentions of Search and Experience Products. *Journal of Interactive Marketing*, *27*, 226-235.

Koh, N. S., Hu, N., & Clemons, E. K. (2010). Do online reviews reflect a product's true perceived quality? An investigation of online movie reviews across cultures. *Electronic Commerce Research and Applications*, *9*, 374-385. http://dx.doi.org/10.1016/j.elerap.2010.04.001

Li, X., & Hitt, L. M. (2008). Self-selection and information role of online product reviews. *Information Systems Research*, *19*, 456-474.

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

Nielsen. (2013). *Under the influence: Consumer trust in advertising.* Retrieved from http://www.nielsen.com/us/en/newswire/2013/under-the-influence-consumer-trust-in-advertising.html

Pew Internet & American Life Project History. (2013). Retrieved from http://pewinternet.org/Static-Pages/About-Us/Project-History.aspx

Punj, G. N. (2012). Do consumers who conduct online research also post online reviews? A model of the relationship between online research and review posting behavior. *Marketing Letters, 24,* 97-108.

Riloff, E., & Wiebe, J. (2003). Learning extraction patterns for subjective expressions. In Proceedings of the 2003 conference on empirical methods in natural language processing. *Association for Computational Linguistics*, 105-112.

Sundaram, D. S., Mitra, K., & Webster, C. (1998). Word-of-mouth communications: A motivational analysis. *Advances in Consumer Research, 25*, 527-531.

Trusov, M., Bodapati, A., & Bucklin, R. E. (2009). Determining influential users in internet social networks.

Wang, X., Yu, C., & Wei, Y. (2012). Social Media Peer Communication and Impacts on Purchase Intentions: A Consumer Socialization Framework. *Journal of Interactive Marketing, 26*, 198-208. http://dx.doi.org/10.1016/j.intmar.2011.11.004

Wang, Z. (2010). Anonymity, social image, and the competition for volunteers: A case study of the online market for reviews. *The BE Journal of Economic Analysis & Policy, 10*.

Wiebe, J., & Riloff, E. (2005). Creating subjective and objective sentence classifiers from unannotated texts. In Computional Linguistics and Intelligent Text Processing (pp. 486-497).

Zhang, J. Q., Craciun, G., & Shin, D. (2010). When does electronic word-of-mouth matter? A study of consumer product reviews. *Journal of Business Research, 63*, 1336-1341.

http://dx.doi.org/10.1016/j.jbusres.2009.12.011

Zhu, F., & Zhang, X. (2010). Impact of online consumer reviews on sales: The moderating role of product and consumer characteristics. *Journal of Marketing*, *74*, 133-148.