Textual vs. numerical rating scales in e-commerce reviews

ERASMUS UNIVERSITEIT ROTTERDAM

Faculty of Economics of Business

Marketing

Name: J.P. van Unnik

Student number: 356956

E-mail address: j.vanunnik@student.eur.nl

Study: International Bachelor Economics and Business Economics (IBEB)

Supervisor: Drs. Havranek

Thesis: Bachelor

Table of contents

Page 1 Table of contents

Page 2 Introduction

Page 5 Theoretical Framework

Page 9 Methodology

Page 11 Results

Page 14 Conclusion

Page 14 Limitations

Page 15 Theoretical Implications

Page 15 Future Research

Page 16 Literature List

Page 18 Appendix

Introduction

Nowadays people use the internet to satisfy their needs for a wide range of products and services. Online sales (e-commerce) has boomed over the last years. Many existing companies established online presence and new ‘online-only’ shops opened their doors. According to Euromonitor, global e-commerce is currently becoming an trillion dollar industry (figure 1). Although growth rates are declining, they have been and still are strong (15%<). This has led to an 8.4% share of retail trade for online retail in Europe (Statista, 2016), making online shopping an important field of interest to many entrepreneurs and researchers. There is a strict distinction between online and offline shopping, when only taking into account on what platform the sale is conducted. Some products are better suited to be sold online and vice-versa. A 2005 study by Levin, Levin and Weller revealed that when (potential) customers prefer a large selection and quick purchase experience, online shopping is preferred. For products where personal service and the ability to feel, touch and handle is considered to be important (e.g. clothing), respondents indicated to prefer offline shopping (Levin, Levin & Weller, 2005). For many practitioners it is therefore a challenge to offer the browsing and purchasing experience (potential) customers demand. This has led to a replication of offline attributes and experiences in the online market. Customers can for example return purchases and claim refunds after feeling, touching and handling products at home. Online personal assistance through customer service has flourished as companies continue to develop their competitive advantages by expanding and innovating their services. And maybe most of all, online word of mouth (WOM) has established unprecedented importance as the internet tremendously increased interconnectivity of individuals. The average customer is able to share their experiences and stories with every user of the world wide web. Nowadays a big part of the online WOM is communicated through a common format: a user review. These often consist of a product or business rating, accompanied by a small description. This familiar format enables internet users to quickly gather product and business information and the open architecture of the internet more or less assures the reliability of the average review by providing large quantities of transparent data.

The online review sentiment is mainly positive. When browsing the biggest online retailer, which is Amazon according to Alexa.com, the products with a significant amount of reviews will often show a J-shaped review distribution (figure 2). In this distribution the amount of reviews for 2- to 5-star reviews is exponentially increasing. The amount of 1-star reviews is often breaking this trend by being larger in count than the 2-star reviews.

This reveals a potential information problem, as the average score of the average product within a product range, is often not reflected by the average score of an uniform scale (maximum score * 0,5). Ratings are often labeled, for example from “awful” to “excellent” (figure 3). If the average product of a retailer is rated with 4 out of 5 stars, one could wonder if their products are “good” on average, the shopping experience with the retailer had a positive influence on ratings or the textual and numerical scales are not internally valid.

Pre-purchase evaluations by (potential) customers studies have found that reviews with extreme scores have more information value. Especially reviews with low(er) ratings are inspected more thoroughly by people. (Lelis & Howes, 2011). The given score by a review is just one of its attributes. Users often also have the ability to provide a textual elaboration. This content can be analysed using Natural Language Processing (NLP). A recent study found that the content of 4 and 5 star reviews contain significantly more (positive) emotional content than 1 star reviews. Surprisingly, the length of these ‘emotional reviews’ is shorter than 2 and 3 star reviews and provide less substantive and objective information. This would suggest that less extreme reviews have more non-emotional information value to users. (Ullah, Amblee, Kim & Lee, 2015). This would mean that a very valuable range of information is consistently overlooked by consumers.

Many studies have found that user reviews are generally positive (Chevalier & Mayzlin, 2003), (Hu, Pavlou & Zhang, 2009), (Lelis & Howes, 2011), and people are more likely to share a positive purchase experience through a review (Ullah, Amblee, Kim & Lee, 2015). Other studies suggest there might also be a purchasing and under-reporting bias. For example, the fact that people tend to buy books that they think they’re going to like, might lead to a positive voting bias in the population (Chevalier & Mayzlin, 2003).

In addition to this purchasing bias, the under-reporting bias is also a suggested cause of the typical review distribution (Hu, Pavlou & Zhang, 2009). This bias indicates that the more extreme one's opinion is, the higher the chance he or she will take the effort of reviewing a product or service, leading to extreme distributions.

These suggested explanations all originate from certain pre-conditions that might exists before an actual review is completed. Both the purchasing bias and under-reporting bias can explain how a group of internet users with non-normal distributed opinions conduct a review, leading to the review distributions we observe today. But none of these studies and theories look at the properties of common online reviews themselves. There are different ways to ask a customer feedback on their recent purchase. Companies differ between asking for a numerical score and/or textual attitude. Asking for a customer' attitude through an ordinal scale may influence the customer' perception and interpretation of scales that are commonly used. Therefore this research asks:

Do online retail post-purchase evaluations differ between textual and numerical review scales?

Exploring and answering this research question fills a gap in recent academic e-commerce literature. The study also aims to help both e-commerce practitioners and third party review platforms with choosing a post-purchase evaluation format that suits their goals and needs. To answer the research question a theoretical framework will first be established. This framework will provide the reader with an understanding of what post-purchase evaluation methods are generally used by e-commerce practitioners and theories of recent studies on J-shaped review distributions and their impact and causes. The second part of the framework will distinguish between textual and numerical review methods. This will be the fundament of the following A/B-type experiment that will assess textual and numerical review scales and the critical evaluation of this experiment. The thesis concludes with the overall findings and recommendations on this subject.

Theoretical Framework

The purchase- and under-reporting bias are graphically displayed over a J-shaped distribution in figure 4. Here a normal distribution is compared to a typical J-shaped review distribution. It is suggested that the gaps around median ratings can be explained by the underreporting bias and for extreme ratings by the purchasing bias. Another suggestion that could explain the particular distribution of reviews is that the scale might not be uniformly distributed. This lies at the core of this research, as it is very important in the classification of interval and ordinal variables. The options of a reivew that are given to a respondent are ordered. Two stars is greater or 'better' than one star, just like "good" is a higher value than "average". In the latter example it is not possible to deduct a value from the given options. Therefore it is impossible to state that "good" is twice as big as "average". However, when assigning numerical elements to the review options (e.g. 1 stars through 5 stars) a new dimension is added. It is still hard to claim that 4 stars is twice as big as 2 stars, since the scale is bound to the highest value. Because in that line of reasoning there can't be a value twice as big as 4 stars. However, the differences between ratings are getting more significance, as going from 3 stars to 2 stars, is the same shift in absolute terms as when going from 3 stars to 4 stars. This is typical for an interval variable, as the values themselves have no meaning, but the differences between them do.

This does not apply to textual scale options. When looking at the Amazon example (figure 3), "fair" is not necessarily exactly in-between "awful" and "excellent". Here there can be a difference in customer interpretation. This is supported by a classical experiment that asked respondents to assign weightings to the adjectives on a seven-point scale (Lodge, 1981). It was found that going moving from "So-So" (mode value) to "Bad" or "Good" actually had a different magnitude and that people are more easily inclined to assign higher values.

So far we have discussed examples and suggested causes of online reviews with a J-Shaped distribution. These are all observations of online available reviews. Although the underreporting and purchasing bias provide some possible causes for the non-normal distribution, they might not fully explain the observations. Each individual might attach different values to review scales, as there are different ways to approach them.

If the scale is seen as the relative performance against all businesses in a sector, the median rating would reflect an average business. However, when it is seen as a testing scale based on the amount or degree of errors incurred, as often used in schools and universities, the highest score would mean that simply nothing is or went wrong. Just like an "A" or "10" can be the highest achievable test result, when an examinee provides no faulty responses. In light of this it is all a matter of perspective how one completes a post-purchase review. In order to assist consumers during the rating process, companies often provide textual elaboration to star ratings, which can guide them in attaching a more or less uniform value to the ratings. There are differences in how companies describe different ratings, but the majority of international webshops use a scale where the median is "average" or something similar (figure 6).

On the statistical side of things, studies approach online reviews in different ways. They discuss the proper way of analysing these reviews. For example, whether ordinal or interval statistics are appropriate for a likert scale (Göb, McCollin & Ramalhoto, 2007). There is a lot of controversy on this matter. Many studies recode the adjectives used in scales into numerical scores and apply interval statistics, even without proper reasoning (Jamieson, 2004). It has been and still is a debate on how to treat the data,, but nearly everyone agrees on the use of ordinal statistics (Knapp, 1989). For example, median and modes are accepted, but the use of averages and standard deviations are doubtful.

This discussion rose after the introduction of the Likert scale. Many studies refer back to older attitude measurement frameworks, of which SERVQUAL (Service Quality) is used very often. These methods use elaborate questionnaires, for example to analyse the gap between quality perception and actual experiences. They use Likert scales to measure an individual's attitude towards a subject. The internet served as a catalyzer for such instruments, by facilitating distribution and ease. Today it is common practice to send feedback requests from webshops. Even though the average review process is not very similar to a 22 items questionnaire (as suggested by Parasuraman et al., 1985, 1988), the analytical reasoning behind it can still apply. We are still facing the same methodological issues as mentioned before. There are many critical evaluations of different ways to analyze responses generated from Likert-like scales.

One of these evaluations, a paper by Göb, McCollin and Ramalhoto, sum up five criteria that should apply to appropriate attitude measuring (fig. 5). These five criteria will be used to evaluate different review methods and results:

● Longitudinal Consistency	respondents give consistent ratings over time
● Longitudinal Comparability	respondent ratings can be compared over time
● Internal Consistency	respondents give consistent ratings
● Interpersonal Comparability	all ratings measure the same thing
● Plausibility	the data received or used is credible

The paper suggests that when the variables are treated as ordinal variables, the analysis is less restricted in terms of where the data originates from. This is not the case for interval variables and the transformation/recoding from an ordinal variable to an interval variable. This often leads to misinterpretations of the data and results, such as “We’ve increased customer satisfaction by 150% in one year” - merely based average rating mutations. This is because the distances between options on an ordinal scale are not assumed equal. Therefore it is striking that many webshops display averages and other similar interval-variable statistics to summarize their ratings. This should be taken into careful consideration when analyzing the results of this research.

This research wants to dig deeper in this natural perception of different scales by online customers. All of the above raises the idea that when a large group of customers is asked to review a purchase through a numerical scale, it is likely that their interpretations of this scale are less consistent as compared to textual scales that provide a subjective understanding. To find whether there is difference between these interval and ordinal scales, this research want to see if the distribution of star-only scales differs from textual scales, which leads to the following hypothesis.

H_{01}: *The distribution of responses from an five points textual review scale is different from the distribution of responses from an five points numerical review scale.*

The majority of online reviews are primarily numerically scaled and they have a tendency to be extremely distributed and skewed towards positive ratings. As the majority of textual scoring options set the median score as average and may influence interpretation of the scale, it might follow that their responses are less extremely distributed, following:

H_{02}: *The distribution of responses from an five points textual review scale are less extremely distributed than the distribution of responses from an five points numerical review scale.*

Methodology

In order to investigate these hypotheses an experiment is conducted. This experiment aims to find if people interpret numerical and textual ratings differently. Besides showing that people might interpret two common label sets of the same scale differently, this research also wants to test this in the most realistic setting as possible. Since the research question is based on observations from online ratings of e-commerce businesses, the experiment will be conducted on customers of a webshop. A Dutch online retailer was found willing to execute the experiment on over 2200 of their e-mail newsletter subscribers. This retailer sells and ships around 600 different national and international beers to consumers in The Netherlands and Belgium. They operate in the full spectrum of an online retailer. Their products cover a small part of the whole e-commerce market, but they operate using common methods and platforms.

The business uses Mailchimp as their third party digital newsletter platform. This provides to option to send emails to two random generated recipient groups, who are served a different version of the newsletter. The platform can track who opens the emails and clicks on its content, which will take the recipient to the website of the retailer. Furthermore the system receives information about recipient behaviour on the website, for example to track the sales generated by the newsletter. The rating given by a recipient is collected on the sending and receiving end, where the e-mail client can 'send' the recipient to the webshop after clicking an URL. When one chooses a rating, they are routed to the webshop via a proxy of Mailchimp. Every URL in the newsletter is unique for each link and recipient. This way the system can deduct the clicked url on a per recipient basis. On the receiving end (the webshop), all recipients load the same page with one of ten URL suffixes. Every review option has its own URL suffix. This 'page load' with the corresponding suffix is registered by a script on the site and reported to Google Analytics. This yields a dataset where we can see the aggregate page loads for each review options. The first method is regarded as the most reliable as it is on a per user bases and can detect multiple clicks/reviews per user. Therefore this will be used as primary data source and Google Analytics as secondary data source to verify the results.

Little is known about the demographic composition of the recipients, but their affinity with the shop is very high: over 51% of the recipients opens the newsletter and over 14% clicks on its content (figure 8). This is at least 3 times higher than the industry average for e-commerce newsletters (Mailchimp, 2016). Nearly all of their newsletter subscribers signed up for regular mailings when completing a purchase or through active subscription on the website. The retailer does have one brick and mortar store, but nearly all subscriptions are generated online. In addition, not the company name, but the website address is used in the newsletter subject and content. This is to make sure it is referring to the online experience of recipients. Since the aim of the experiment is to test the formulation of the review options ceteris paribus, both e-mails are identical, except for the text within the actionable buttons of the newsletter (figure 10). The design of the newsletter is also in line with templates and designs used for prior mailings.

This method of conducting the experiment has a strong external validity as it is almost indistinguishable from actual review requests that are sent out by companies. Only a small disclaimer below the given review options refer to that the customer input may be used for research purposes. The aforementioned interpersonal comparability (figure 5) remains a potential issue, as the two scales differ and therefore may measure different things. This impacts all statistical reasonings resulting from comparing the two samples on a fundamental level. Yet it is closely related to the hypothesis, as we do want to see if the two scales actually yield different distributions. Therefore it is assumed that those who use either scale (e.g. webshops), attribute the same underlying values to both of them and it is only the respondent/reviewee who might hold a different interpretation between the scales (which would confirm the hypothesis).

Results

Exactly 200 unique responses of recipients have been collected in the primary data set (figure 11). The secondary data contains over 300 visits. Since the secondary set can contain multiple visits of a recipient and the number is not below the 200 expected visits resulting from the primary dataset, the data is assumed to be correct. The primary dataset presented the following distributions of ratings:

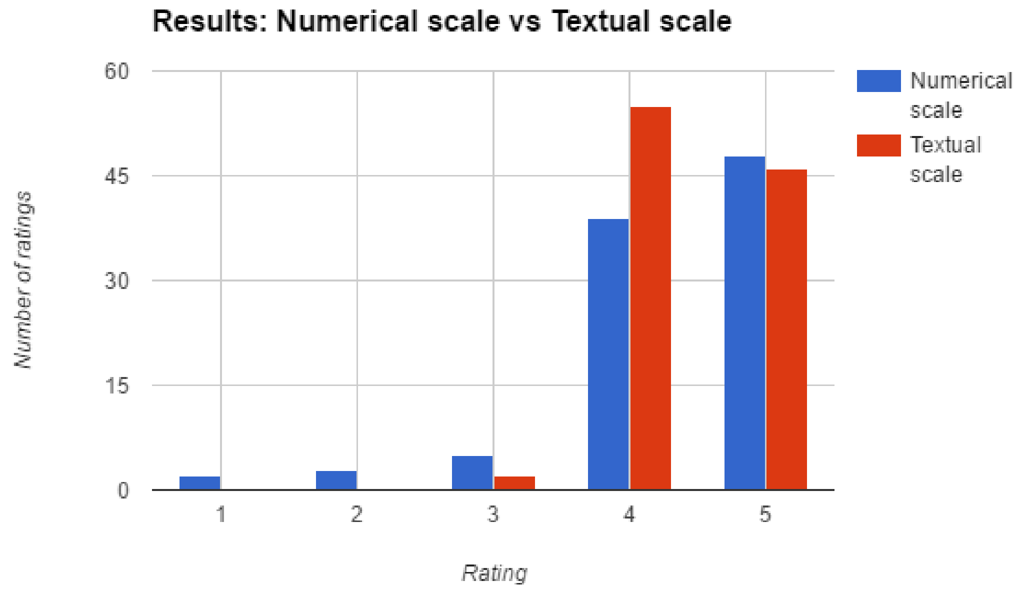


Figure 12: distributions of experiment ratings per mailing and scale option

The textual scale recorded ‘good’ as the most rated option and the numerical scale recorded 5 stars as the most rated option. (figure 12). 94% (118) of all ratings are positive (4 stars, 5 stars, good or very good).

The goal is to analyze the two different distributions of the textual and numerical ratings. Visually there are some differences. The textual scale received no negative ratings (bad, very bad, 1 star or 2 stars), where the numerical did. The numerical scale seems more evenly distributed over the scale and received mostly ‘5 star’ ratings, in contrary to the textual scale, which mostly received “good” ratings.

The answer whether the two distributions can be assumed to be different, can be provided after further analyzing the results using statistical analysis. Both parametric and nonparametric statistics will be used, since both may have their own explanatory power. The conclusion and discussion of this thesis will address the applicability of the resulting statistics and discuss the validity of the used tests.

The conducted parametric tests are chi-square metrics, based on expected values and the means of the distributions. The cross-tabulations of both scales against their respective ratings provided a Pearson Chi-Square statistic indicating a weak association (p = 0.064) between the type of scales and ratings given (figure 13). As the count of some expected values are below 5, the Chi-Square statistic tends to overvalue the significance for smaller numbers in its calculations. Therefore the Pearson Chi-Square was recalculated for all positive rating, as these make up the majority of the data set for both scales. This resulted in a P-value of 0.188, indicating no significant association between the type of scales and the ratings. (figure 14).

The Pearson Chi-Square test has a limitation in this case, since it doesn't take the ordering of ordinal variables into account. Therefore the Linear-by-Linear value is also taken into account. It is based of the Pearson Chi-Square measure but holds ordering into account and is less sensitive for smaller numbers (Agresti, 1996). The linear-by-linear association of the two scales has a significance of 0.290 and therefore also shows no significant association between the type of scale and the ratings.

As mentioned before it is disputable if we can use these parametric tests for the ordinal variables in this experiment, as they are based on means of the samples. Therefore also nonparametric tests have been conducted on the data. Because the Mann-Whitney U test is nonparametric and doesn't assume a normal distribution, it was used first. The Mann-Whitney U test was run to determine if there were differences in ratings between numerical and textual scales.

Distributions of the ratings for numerical and textual scales were not similar, as assessed by visual inspection. There was no statistically significantly difference in engagement scores between numerical and textual scales (U = 4995.500, z =0, p = 1) using an exact sampling distribution for U (Dineen & Blakesley, 1973).

Since the outcome of the Mann-Whitney U test suggested completely equal distributions because both mean ranks of the samples are equal (110.5), another nonparametric test was conducted. The Kolmogorov-Smirnov test looks for more deviations from the same null hypothesis as the Mann-Whitney U test . Therefore it focusses less on differences between the median of two samples but is more sensitive to the differences between distributions (Lehmann, 2006). According to the Kolmogorov-Smirnov test, distributions of ratings for numerical and textual scales were not similar, as assessed by visual inspection. There was no statistically significantly difference in engagement scores between numerical and textual scales (z = 0,591, p = 0.875).

Since the majority of all ratings were positive, the nonparametric tests were again repeated for only the positive ratings (good, very good, 4 stars and 5 stars). This produced a less but still significant relation between the two scale distributions. However, since the neutral and negative ‘outliers’ were no data entry errors, measurement errors or could neither be classified as ‘genuinely unusual values’, these results are not deemed valid.

The results of the statistical tests on the data are summarized below:

Test	Relations
Pearson Chi-Square	Weak relation between type of scale and ratings
Pearson Chi-Square* (positive ratings)	No relation between type of scale and ratings
Linear-by-linear association	No relation between type of scale and ratings
Mann-Whitney U test	No difference between type of scale and ratings
Kolmogorov-Smirnov Two Sample	No relation between type of scale and ratings

Figure 15: conclusions drawn from the five different test statistics

Conclusion

This thesis provides an elaborate review on the distributions of online review scales. The majority of online reviews have a J-shaped distribution and a positive tendency. According to literature, this could be caused by a purchasing- and underreporting bias under consumers. To complement existing literature it was investigated whether online post-purchase evaluations differ between textual and numerical review scales. We expected that customers interpret these scales differently, as they might assign different values to the options of both scales.

This was researched by distributing two two review request with a textual and numerical scale to customers of an online beer retailer. Exactly 200 responses were collected from the email newsletter. The observations of current literature were confirmed by the data. The responses showed a positive tendency and the distribution matches J-shaped properties. However, no significant differences between the data distributions collected from the textual and numerical review scale were observed. The statistical tests that did suggest some degree of difference between the scales, are limited in their interpretation. They would only apply if we assume that the options of the scales are linearly distributed, which is not supported by literature for these types of scales. Therefore we reject our first and second hypothesis, as the distributions of the numerical and textual scale responses are similar.

Limitations

The research was conducted on a very specific part of the online retail market. Furthermore the research was not performed within an actual post-purchase situation, as the respondents were newsletter subscribers and did not necessarily completed a purchase in recent history. This should be taken into account when applying the conclusion on online retail as a whole. However, the results of the research does confirm and complement existing literature, as no contradictory results were found.

The 200 responses were assumed to be a strong starting point for the analysis. But as we have observed during the statistical processing of the data, the nominal nature of the variables limits the significance tremendously. Both scales only had 5 options and therefore had a very limited amount of possible median values for nonparametric testing.

This is probably why the parametric tests yielded a stronger difference between the distributions of the textual and numerical scales, as the variables are considered continuous in such tests.

Theoretical Implications

The positive tendency and J-shaped distributions of online reviews provide both opportunities for businesses. The underreporting bias can be used to improve ratings, by actively engaging customers that are expected to have had a positive purchasing experience. This degree of control that companies have over their own ratings, provides another reason for the existence of independent third-parties that provide ratings about companies. The underreporting bias means a lack of valuable information in the middle segment (e.g. 2 to 4 stars) of the common review scale. Third parties can provide this information by collecting and segmenting this information from customer feedback along the review scale.

If there is a difference in interpretation between textual and numerical scales, this can aid third parties by guiding their reviewers in attaching a desirable score to feedback. This of course also applies to companies who want to format their reviews in such a way, that it will yield them the highest score. Of course, such a decision should be supported by a research that does indicate a significant difference between different review scales.

Future research

This research provides a fundament of relevant literature and statistical methods for processing online review data. It might be limited due to nonparametric testing on a five point scale. Therefore, future research might yield a significant difference between textual and numerical scales, if the amount of respondents is higher. So repeating this study with a greater amount of respondents has the potential of yielding significant results. This research could also be repeated using different types of scales or labels, as only one set has been tested. If no difference in interpretation is found between the textual and numerical review scales, using other scales and labels might have an effect. All of the above would provide a greater understanding of how customers interpret online review scales and how these are suitable for companies in different situations.

Literature List

Alexa (2016) http://www.alexa.com/topsites (fetched on 17-05-2016)

Agresti, A., & Kateri, M. (2011). Categorical data analysis. Springer Berlin Heidelberg.

Chatterjee, P. (2001). Online reviews: do consumers use them?. 129-134.

Dawson, L., Minocha, S., & Petre, M. (2003). Exploring the total customer experience in e-commerce environments. Proceedings of the IADIS International Conference e-Society.

Dinneen, L. C., & Blakesley, B. C. (1973). Algorithm AS 62: A generator for the sampling distribution of the Mann-Whitney U statistic. Journal of the Royal Statistical Society. Series C (Applied Statistics), 22(2), 269-273.

Duan, W., Gu, B., & Whinston, A. B. (2008). Do online reviews matter?—An empirical investigation of panel data. Decision support systems, 45(4), 1007-1016.

Exploiting Ordinality in Predicting Star Reviews. Virani, Cameron (http://www.cs.ubc.ca/~carenini/TEACHING/CPSC503-16/PROJECTS-14/final_paper-ChrisAlim.pdf - retrieved on 17-05-2016)

Göb, R., McCollin, C., & Ramalhoto, M. F. (2007). Ordinal methodology in the analysis of Likert scales. Quality & Quantity, 41(5), 601-626.

Hu, N., Zhang, J., & Pavlou, P. A. (2009). Overcoming the J-shaped distribution of product reviews. Communications of the ACM, 52(10), 144-147.

Knapp, T. R. (1990). Treating ordinal scales as interval scales: an attempt to resolve the controversy. Nursing research, 39(2), 121-123.

Lehmann, E. L., & D'Abrera, H. J. (2006). Nonparametrics: statistical methods based on ranks (p. 39). New York: Springer.

Lelis, S., & Howes, A. (2011). Informing decisions: how people use online rating information to make choices. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM.

Levin, A. M., Levin, I. P., & Weller, J. A. (2005). A multi-attribute analysis of preferences for online and offline shopping: Differences across products, consumers, and shopping stages. Journal of Electronic Commerce Research, 6(4), 281.

Lodge, M. (1981). Magnitude scaling: quantitative measurement of opinions.

Mailchimp (2016). Email Marketing Benchmarks http://mailchimp.com/resources/research/email-marketing-benchmarks/ fetched on 12-06-2016

Mayzlin, D., & Chevalier, J. A. (2003). The effect of word of mouth on sales: Online book reviews. Yale School of Management Working Papers. Yale School of Management.

Minocha, S., Dawson, L., Roberts, D., & Petre, M. (2004). E-SEQUAL: A Customer-centred Approach to Providing Value in E-commerce Environments. Department of Computing Faculty of Mathematics and Computing. The Open University Walton Hall, United Kingdom.

Petre, M., Minocha, S., & Roberts, D. (2006). Usability beyond the website: an empirically-grounded e-commerce evaluation instrument for the total customer experience. *Behaviour & Information Technology*, *25*(2), 189-203.

Resnick, P., & Zeckhauser, R. (2002). Trust among strangers in internet transactions: Empirical analysis of ebay's reputation system. The Economics of the Internet and E-commerce, 11(2), 23-25.

Shankar, V., Smith, A. K., & Rangaswamy, A. (2003). Customer satisfaction and loyalty in online and offline environments. International journal of research in marketing, 20(2), 153-175.

Statista (2016). http://www.statista.com/statistics/281241/online-share-of-retail-trade-in-european-countries/ (fetched on 17-05-2016)

Ullah, R., Amblee, N., Kim, W., & Lee, H. (2016). From valence to emotions: Exploring the distribution of emotions in online product reviews. Decision Support Systems, 81, 41-53.

Winkelmann, A., Herwig, S., Poeppelbuss, J., Tiebe, D., & Becker, J. (2009). Discussion of functional design options for online rating systems: A state-of-the-art analysis. ECIS.

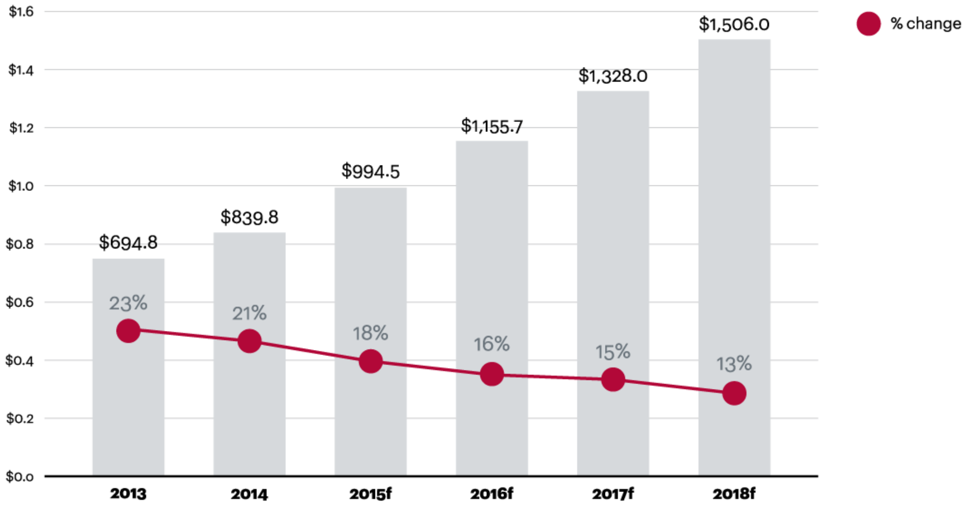
Appendix

Figure 2

Global retail sales are set to increase, although the rate of growth will slow

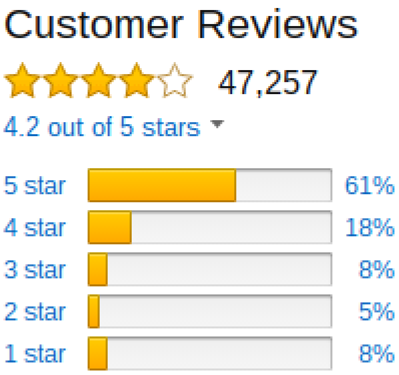
Global e-commerce sales

(US$ billions)



Source: Euromonitor

Figure 1: data collected by A.T. Kearney from Euromonitor on global e-commerce trends and predicitons



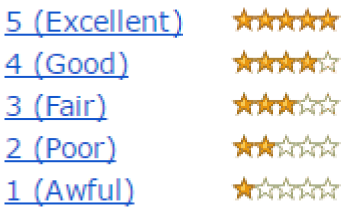
Customer Reviews

★★★★☆ 47,257

4.2 out of 5 stars

5 star	61%
4 star	18%
3 star	8%
2 star	5%
1 star	8%

Figure 2: J-shaped review distribution on Amazon. http://www.amazon.com/Amazon-Fire-7-Inch-Tablet-8GB/product-reviews/B00TSUGXKE/ - retrieved on 09-05-2016



5 (Excellent)	★★★★★
4 (Good)	★★★★☆
3 (Fair)	★★★☆☆
2 (Poor)	★★☆☆☆
1 (Awful)	★☆☆☆☆

Figure 3: Rating explanation of Amazon.

Retrieved from personal e-mail sent on 15-08-14

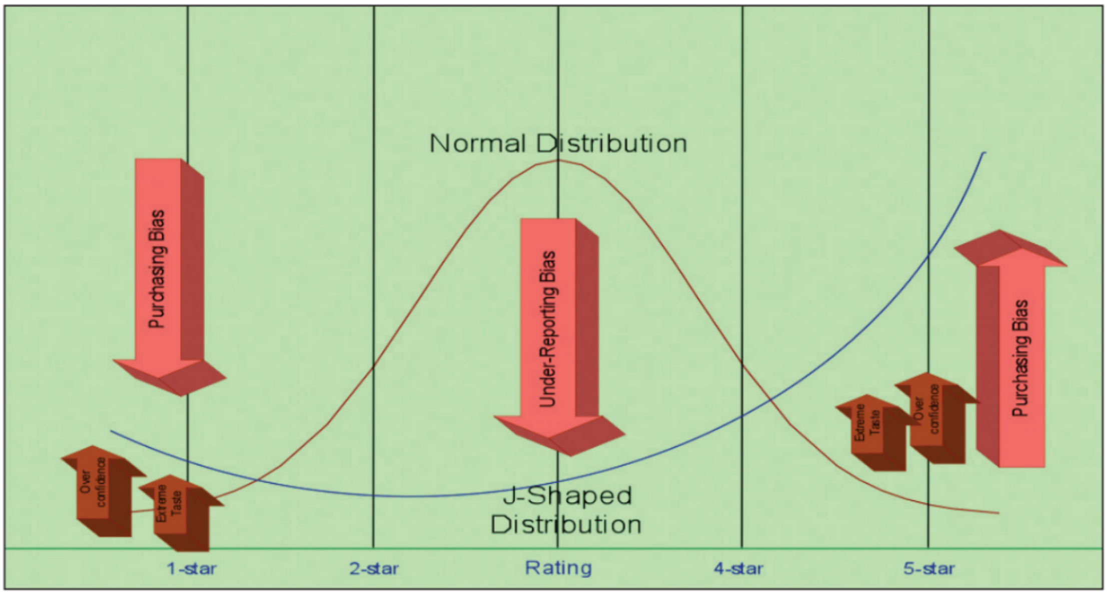
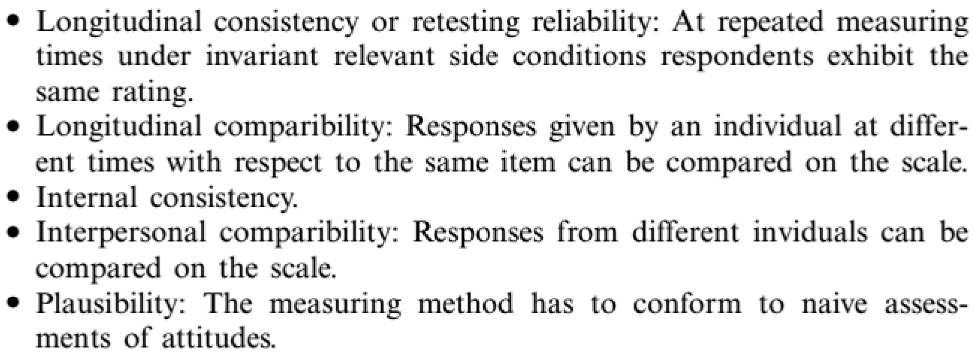


Figure 4: Purchasing- and Under-reporting Bias

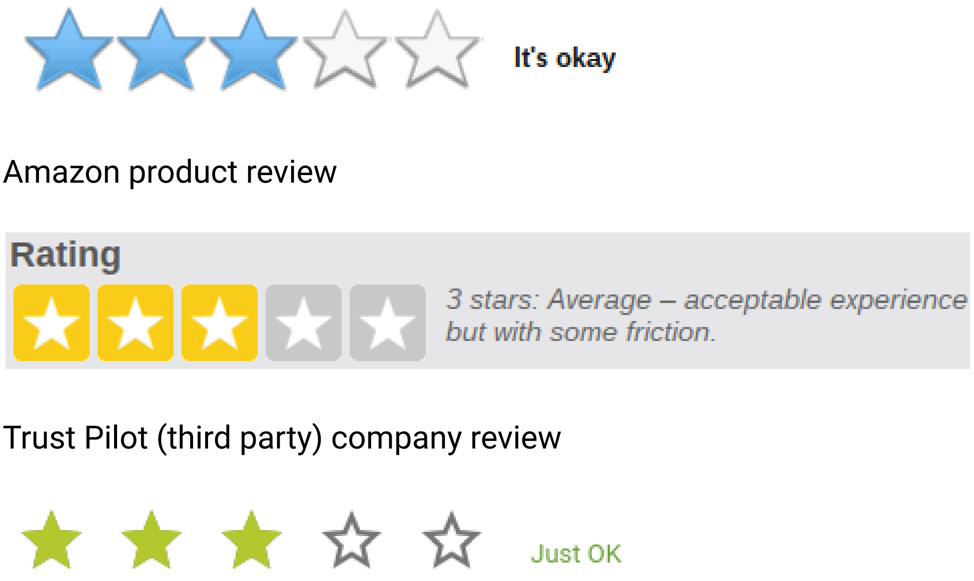
(Hu, Pavlou & Zhang, Overcoming the J-Shaped Distribution of Product Reviews (2009)



- Longitudinal consistency or retesting reliability: At repeated measuring times under invariant relevant side conditions respondents exhibit the same rating.
- Longitudinal comparibility: Responses given by an individual at different times with respect to the same item can be compared on the scale.
- Internal consistency.
- Interpersonal comparibility: Responses from different invidual s can be compared on the scale.
- Plausibility: The measuring method has to conform to naive assessments of attitudes.

Figure 5: five attitude measuring criteria.

Ordinal methodology in the analysis of Likert Scales. Göb, McCollin, Ramahoto (2007)

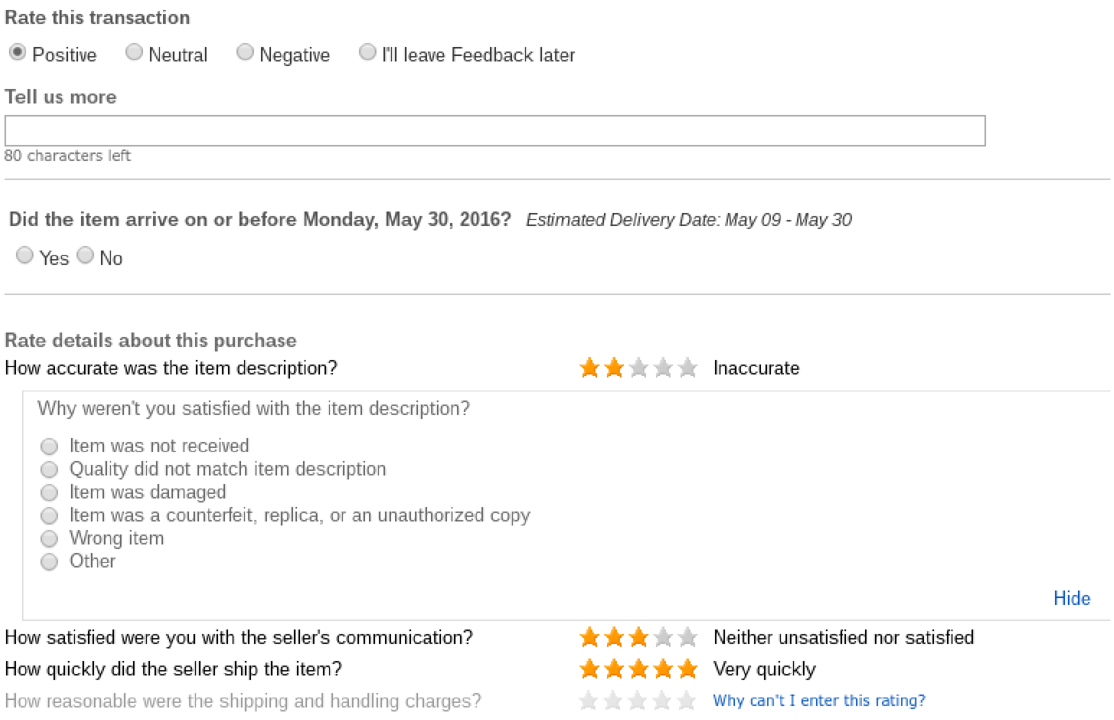

It's okay
Rating
3 stars: Average – acceptable experience but with some friction.
Just OK

Amazon product review

Trust Pilot (third party) company review

Google Play Store review

Figure 6: Textually assisted star ratings



Rate this transaction

● Positive ○ Neutral ○ Negative ○ I'll leave Feedback later

Tell us more

80 characters left

Did the item arrive on or before Monday, May 30, 2016? *Estimated Delivery Date: May 09 - May 30*

○ Yes ○ No

Rate details about this purchase

How accurate was the item description? ★★☆☆☆ Inaccurate

Why weren't you satisfied with the item description?

○ Item was not received
○ Quality did not match item description
○ Item was damaged
○ Item was a counterfeit, replica, or an unauthorized copy
○ Wrong item
○ Other

Hide

How satisfied were you with the seller's communication? ★★★☆☆ Neither unsatisfied nor satisfied

How quickly did the seller ship the item? ★★★★★ Very quickly

How reasonable were the shipping and handling charges? ☆☆☆☆☆ Why can't I enter this rating?

Figure 7: eBay post-purchase rating process

	Open rate	Click rate
Used subscriber list	51.06%	14.4%
e-commerce industry average	16.77%	2.46%

Figure 8: average rates before the experiment (last 5 newsletter to all subscribers) and industry average according to Mailchimp (as of 02-06-2016).

	E-mails sent	Delivery rate	Open rate	Click rate
Numerical (stars)	1113	96.6%	48.1%	8.9%
Textual (text)	1112	96.3%	47.5%	9.7%
Difference	*1*	*0.3%*	*0.6%*	*1.0%*

Figure 9: newsletter experiment send data (as of 15-06-2016)

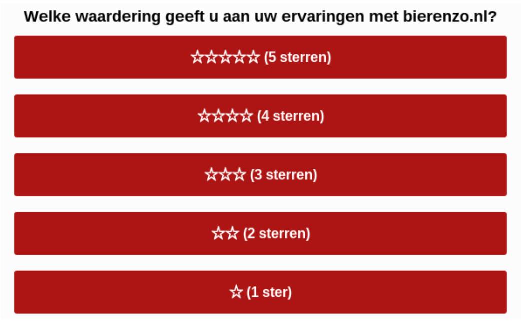
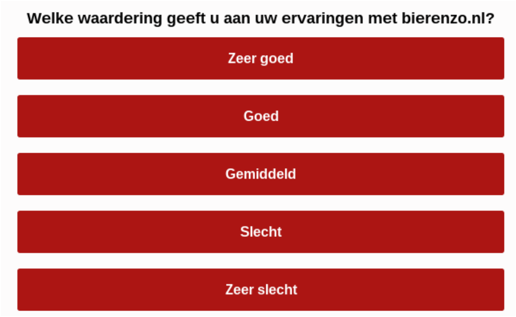
Email A (numerical)	Email B (textual)
 Welke waardering geeft u aan uw ervaringen met bierenzo.nl? ☆☆☆☆☆ (5 sterren) ☆☆☆☆ (4 sterren) ☆☆☆ (3 sterren) ☆☆ (2 sterren) ☆ (1 ster)	 Welke waardering geeft u aan uw ervaringen met bierenzo.nl?

Figure 10: the two different newsletter email sent to the experiment recipients.

	Numerical scale	Textual scale	Total
5 stars / ‘very good’	48	46	94
4 stars / ‘good’	39	55	94
3 stars / ‘average’	5	2	7
2 stars / ‘bad’	3	0	3
1 star / ‘very bad’	2	0	2
Total	*97*	*103*	*200*

Figure 11: experiment ratings per mailing and scale option

	Numerical scale	Textual scale	Total
Mean*	4.33	4.43	4.38
Median	4	4	4
Mode	5	4	4

Figure 12: Descriptive frequency statistics of the two scales

*these are ordinal variables. Mean value might not be suitable statistic.

All ratings	Test value	Significance
N	200	
Pearson Chi-Square*	8.880	0.064
Linear-by-linear association	1.119	0.290
Mann-Whitney Z-value	0	1.00
Kolmogorov-Smirnov Z-value	0.591	0.875

Figure 13: Test statistics and significance comparing the distributions of the two scales .

*some expected counts are below 5

4 stars, 5 stars, ‘good’, ‘very good’	Test value	Significance
N	188	
Pearson Chi-Square	1.733	0.188

Figure 14: Test statistics and significance comparing the distributions of the two scales, for positive ratings.