In the Eye of the Beholder? Empirically Decomposing Different Economic Implications of the Online Rating Variance[1]

Dominik Gutt
Paderborn University
dominik.gutt@uni-paderborn.de

December 16, 2018

Abstract

The growing body of literature on online ratings has reached a consensus of the positive impact of the average rating and the number of ratings on economic outcomes. Yet, little is known about the economic implication of the online rating variance, and existing studies have presented contradictory results. Therefore, this study examines the impact of the online rating variance on the prices and sales of digital cameras from Amazon.com. The key feature of our study is that we employ and validate a machine learning approach to decompose the online rating variance into a product failure-related and taste-related share. In line with our theoretical foundation, our empirical results highlight that the failure-related variance share has a negative impact on price and sales, while the impact of the taste-related share is positive. Our results highlight a new perspective on the online rating variance that has been largely neglected by prior studies. Sellers can benefit from our results by adjusting their pricing strategy and improving their sales forecasts. Review platforms can facilitate the identification of product failure-related ratings to support the purchasing decision process of customers.

Keywords: Online Rating Variance, Text Mining, Econometrics, User-Generated Social Media.
JEL Classification: M15, M31, O32, D12

[1] The author thanks participants at the 2018 European Conference on Information Systems. The author benefitted from helpful discussions with Hendrik Schmitz and from research assistance by Robin Wulfes. The work was partially supported by the German Research Foundation (DFG) within the Collaborative Research Centre "On-The-Fly Computing" (SFB 901).

1 Introduction

In the era of digitization, user-generated social media content –such as online ratings— are a primary way to acquire online information about goods and services. Online ratings are a driving force behind online (Cabral and Hortaçsu 2010) as well as offline (Anderson and Magruder 2012) customer behavior because they facilitate the evaluation and comparison of the quality of a product or service. In fact, surveys report that 90% of all online purchase decisions are influenced by online ratings (Drewnicki 2013), and they are considered a crucial success factor in Amazon's business (Allen 2016). Unsurprisingly, researchers have been keen on investigating the impact of this information on customer behavior. As a result, the positive effects of the average rating (Luca 2016, Anderson and Magruder 2012, Li and Hitt 2008) and of the number of ratings (Chevalier and Mayzlin 2006, Duan et al. 2008) are well established in the literature. Also, numerous studies have investigated the drivers behind the often-observed J-shape of the rating distribution (Hu et al. 2017, Koh et al. 2010).

Although previous studies generally depict consistent results concerning the impact of the average rating and the number of ratings, the effect of the variance of online ratings on economic outcomes remains insufficiently understood. The variance of online ratings is a metric indicating to what extent customers disagree with their evaluation of the underlying product or service. Even though some empirical studies have investigated the effects of the online rating variance on economic outcomes, these studies have delivered inconclusive evidence. Whereas Clemons et al. (2006) and Lu et al. (2014) found a positive association between the variance of online ratings and sales, Chintagunta et al. (2010), Moon et al. (2010) and Ye et al. (2011) did not find support for such an association; in fact, Ye et al. (2009) even found a negative relationship between the variance and sales.

This lack of knowledge represents a major handicap for sellers and buyers alike. Sellers need to understand the implications of the online rating variance so they can adapt the sales strategies of their products and services. Several studies have highlighted that increasing average ratings enable sellers to increase their sales prices (Lewis and Zervas 2016, Neumann and Gutt 2017, Teubner et al. 2017), but it is unclear how sellers should react to changes in the variance of online ratings. Without knowledge of the effects of the online rating variance on strategic variables such as price and demand, managers cannot be sure if they should change their prices or alter their demand forecast. Interpreting the variance of online ratings also poses a challenge for buyers because they might not be knowledgeable enough to determine whether customer disagreement is a good or a bad sign. Even though disagreement can be beneficial in the way that negative ratings often contain helpful information (Sen and Lerman 2007), substantial disagreement might scare customers who are pleased with buying mainstream products (Sun 2012).

This study, therefore, aims at empirically analyzing the effect of the variance of online ratings on prices and sales by testing propositions from an analytical model (Zimmermann et al. 2017). In this model, the online rating variance is divided into different sources—namely, a taste-related share of the variance and a product failure-related share. This study argues that the taste-related variance share can be beneficial for sales, but the failure-related share can be detrimental to sales. This split hinges on the idea that taste-related variance facilitates future customers' purchase decisions; they can find a product that perfectly matches their taste. Taste-related variance, for instance, can be driven by customer disagreement on the ease of navigating the menu of a digital camera—some might like a simple menu, whereas others might prefer a complex menu with highly adjustable settings. However, customers do agree on their dislike of the product failure component of the rating variance; they do not want their digital camera to malfunction. Thus, we pose the following research question:

Does the source of the variance influence the impact of the variance of online ratings on prices and sales?

To answer our research question, we empirically tested two hypotheses that we adapted from Zimmermann et al. (2017) on a unique panel data set. This data set contains all reviews, monthly amazon sales ranks, and prices for the whole product lifetime of a matched sample of 840 digital cameras from Amazon.com. We employed and validated an unsupervised machine learning approach (Latent Dirichlet Allocation [LDA]) to identify online ratings mentioning product failure in their review texts. This enabled us to calculate the share of the variance related to product failure. We then tested the influence of this share of failure-related online rating variance on the price as well as sales of the respective product. Our empirical results suggest that there is a significant negative effect of the failure-related online rating variance share on prices and sales of digital cameras. However, prices and sales are positively associated with the taste-related variance share.

Thus, our research makes various substantial contributions to the literature that are accompanied by valuable practical implications to sellers and buyers in online rating communities. First, and to the best of our knowledge, we are the first to empirically disentangle the online rating variance into a taste- and a failure-related share and test their effects of both shares on sales and prices. In this way, we contribute to the growing literature on online ratings. We provide additional insights into the nature of the interplay between the online rating variance and sales for which studies have presented inconclusive results thus far. Second, our empirical results provide support to the theoretical model of Zimmermann et al. (2017). Therefore, we support the explanatory power of this model that can potentially serve to reconcile conflicting findings on the effects of the online rating variance. Third, our results highlight important practical implications for customers who rely on user-generated social media to support their purchase decisions. Based on our results, customers should be able to scan online ratings to infer whether the online rating variance of a product is primarily caused by product failure or by taste-related aspects—e.g., by reading the negative reviews. Fourth, sellers can judge the competitive edge of their online rating variance and adjust prices accordingly. In addition, they can incorporate the rating variance into their sales forecasts. Utilizing our presented approach, online rating systems might assist both sellers and customers by facilitating the identification of failure ratings.

The remainder of our work is organized as follows: Section 2 presents the related literature, Section 3 discusses the theoretical background, Section 4 presents the research setup, Section 5 discusses and Section 6 concludes our study.

2 Related Literature

Our study is related to two substreams of the literature on online ratings: the literature (i) on the impact of textual review content, and the literature (ii) on the relationship between the variance of online ratings and economic outcomes.

First, a nascent stream of research has explored different ways of leveraging the information contained in online review texts. Several studies have demonstrated various functions of textual information in online reviews. They can be used for predicting the pricing power of products (Archak et al. 2011), for the inference and surveillance of market structure (Netzer et al. 2012), and for facilitating trade by signaling benevolence and commitment in peer-to-peer markets (Pavlou and Dimoka 2006). Our study contributes to this stream of research by deploying an automated identification of product failures in online review texts and, even more importantly, validating this approach with manual coders.

Second, some studies have already empirically investigated the economic impact of online rating variance in the field and in the laboratory in various contexts. These studies can be classified in four groups as some find a positive, negative, neutral, or mixed relationship, as depicted in Table 1.

Table 1: Signs of the Relationship between Online Rating Variance and Economic Outcomes

Sign of Relationship	Clemons et al. 2006	Bao and Chang 2014	Lu et al. 2014	Kostyra et al. 2016	Chintagunta et al. 2010	Moon et al. 2010	Ye et al. 2011	Ye et al. 2009	Langan et al. 2017	Wu et al. 2013	Minnema et al. 2016	Raguseo et al. 2017	Hut et al. 2017	Sun (2012)	He and Bond 2015	Wang et al. 2015
Focal Product/Service	Craft Beer	Books	Hotels	eBook Readers	Movies	Movies	Hotels	Hotels	Laptops, Digital Cameras	Electronics	Electronics, Furniture	Hotels	Books, DVDs, Videos	Books	Desk Lamps, Flash Drives, Paintings, Music Albums, Ice Cream	Movies, Digital Cameras
Positive	✓	✓	✓													
Neutral				✓	✓	✓	✓									
Negative								✓	✓	✓	✓	✓	✓			
Mixed														✓	✓	✓

Three studies find a positive relationship between the online rating variance and economic outcomes for craft beer (Clemons et al. 2006), books (Bao and Chang 2014), and hotels (Lu et al. 2014). Four papers find a neutral relationship between the variance and economic outcomes for eBook readers in an experimental study (Kostyra et al. 2016), movies (Chintagunta et al. 2010, Moon et al. 2010), and hotels (Ye et al. 2011). Most of the studies (6) actually find a negative relationship between the variance and economic outcomes and these studies electronics (Langan et al. 2017, Wu et al. 2013, Minnema et al. 2016), furniture (Minnema et al. 2016), hotels (Ye et al. 2009, Raguseo et al. 2017), and books, DVDs, and videos (Hut et al. 2017). Finally, three papers find a mixed relationship between the online rating variance and sales for books (Sun 2012), electronics (Wang et al. 2015), desk lamps, flash drives, paintings, music albums, and ice cream (He and Bond 2015).

However, to the best of our knowledge, no study so far has investigated the relationship between different sources of the online rating variance—in our case, taste-related and failure-related variance—and sales. Also, none of the prior studies has investigated product failures as one potential source of the online rating variance. In this way, we contribute to the existing empirical literature on online ratings and the online rating variance by investigating the ambiguous effects between different aspects of the variance on product prices and sales of digital cameras. Empirically, we find that the share of failure-related rating variance leads to lower with sales, whereas the share of taste-related variance leads to higher sales.

3 Theoretical Background

Theoretically, a large share of products traded on online markets can be described by search and experience attributes (Nelson 1970). Search attributes can be inferred prior to purchase by inspecting the product information provided by the manufacturer or seller (Shapiro 1983). For digital cameras, search attributes in product information comprise technical features such as the resolution in megapixels, the item dimensions, and the weight. In contrast to search attributes, experience attributes can hardly be learned before the purchase; usually it occurs solely after purchase. For example, for digital cameras, it is difficult to assess the ease of navigating the camera's menu, the filters or scene modes available, and the feel of the material prior to purchase. The products (craft beer and movies) and services (hotel stays) that have been investigated by prior literature share the common trait of many of their attributes being experience attributes; examples include the taste of a beer, the enjoyment of a movie, or the service quality at a hotel.

Yet, online rating systems have changed the way potential customers can infer experience attributes of products. Online ratings enable potential customers to peer-learn from the digitized word of mouth of other customers (Dellarocas 2003). In the course of this, experience attributes can be transformed into attributes that can be searched and evaluated prior to purchase (Zimmermann et al. 2017). Consequently, potential customers can learn about experience attributes of a particular product without using it (Chen and Xie 2008, Hong et al. 2012, Kwark et al. 2014). For instance, by reading online reviews, potential customers can learn about past customers' experiences concerning the ease of navigating a camera's menu. Some customers might like a simple menu, whereas others might prefer a more complex menu with highly adjustable settings. Customers' disagreements resulting from opposing opinions are thus taste related. Potential customers are able to learn not only about the different taste-related experience attributes associated with digital cameras, but about instances of product failure experienced by other customers. Thus, by reading online reviews, customers can approximate the probability of their product to fail. This product failure probability is a failure-related experience attribute.

Both types of experience attributes—taste related and failure related—are likely to induce additional variance into the online rating variance of a product. Although some reviewers will rate the digital camera with 5 stars because they prefer a rather complicated camera menu with a multitude of setting options and generally like the product, others might dislike digital cameras with complicated menus and give only 2 or even 1 star for the camera, thus inducing further variance—i.e., increasing the extent to which customers disagree—in terms of online ratings. This increases the taste-related variance share in the total online rating variance. Failure-related aspects also induce variance of a product's online ratings because instances in which digital cameras stop working (product failure) can lead customers to a low product rating. This increases the failure-related variance share in the total online rating variance.

The analytical model by Zimmermann et al. (2017) hinges on the different perceptions of both types of customer variances. The key difference between both sources of variance is that all potential customers would agree about their dislike of failure-related variance, regardless of their taste. The taste-related variance share has two effects: casual photographers or elderly people might prefer digital cameras with simple menu navigation, whereas experienced photographers or technically adept customers might dislike cameras having simple menu options.

3.1 Hypothesis Development

In our research environment, users post reviews about their experiences with digital cameras they bought on Amazon.com. Digital cameras exhibit both search (weight, item dimensions, color) and experience (digital filters/scene modes, reaction time, ease of navigation of the camera menu) attributes. It is also possible for digital cameras to exhibit product failure; the autofocus can stop working, buttons can be dysfunctional, or the camera can simply fail to boot.

For retailers, the composition of the total online rating variance of the camera has important implications for the camera's sales strategy. Prior literature has already revealed that changes in certain metrics of online ratings (such as the average rating) influence the sales strategy; for example, they can lead sellers to increase or decrease their prices (Feng et al. 2016, Lewis and Zervas 2016, Neumann and Gutt 2017). All else being equal—holding the total variance and the average rating constant—an increasing share of taste-related aspects in the total variance of a camera's online rating signals high customer disagreement on the camera. More advanced photographers might enjoy the picture quality of a reflex camera, whereas casual photographers might dislike the complexity with regard to the usage of the camera. In response to that, retailers can increase the price of the camera knowing that advanced photographers are willing to pay more than casual users.

Analogous to that, an increasing share of the failure-related aspects in the variance should lead retailers to decrease their prices because customers generally dislike products that malfunction. Thus, as per Zimmermann et al. (2017), we formulate our first hypothesis:

Hypothesis 1: An increase in the share of taste-related (failure-related) online rating variance leads to an increase (decrease) in the product's price, holding total variance and average rating constant.

Yet, an increasing share of taste-related online variance also affects the demand for a focal digital camera. The net effect of an increasing share of taste-related variance is determined by two opposing effects. On the one side, customers who merely ike the digital camera but whose preference for simplicity outweighs the increased price will not buy it. On the other, there are several reasons why an increased share in taste-related variance may increase demand.

First, an increased share of taste related-variance automatically leads to a decreasing share in failure-related variance. Fewer product failures are appreciated by all potential customers, which should, in turn, increase the demand for the product.

Second, due to the increased taste-related variance, customers with a well-matched taste will learn about the product but may not buy it if it has a lower taste-related variance—holding the average rating constant—because they consider it too "mainstream." Moreover, profit-maximizing retailers will not increase the price up to the point where customer loss due to price increases is larger than the attraction of additional customers because of a high taste-related variance.

Third, a higher taste-related variance enables customers to better decide which product to buy. For example, a customer whose taste is matched with a particular product should feel more comfortable with a product that has a larger taste-related variance than with a product having a lower taste-related variance because he is better informed about the product's pros and cons. This should give a potential customer a higher confidence in buying the product (Schlosser 2011).

Fourth, a product with a relatively high taste-related variance might be less suspicious of a product compared to a product with a lower taste-related variance—that consists of relatively positive reviews—when both have the same average rating. The absence of negative reviews for the product with the lower taste variance could undermine the trust in the reviewers by the potential customers. Therefore, we formulate our second hypothesis, as per Zimmerman et al. (2017):

Hypothesis 2: An increase in the share of taste-related (failure-related) online rating variance leads to an increase (decrease) in the demand for the product, holding total variance and average rating constant.

4 Research Setup

Based on our theoretical background, we hypothesize a positive impact of the taste-related variance share on prices as well as sales. Equivalently, we hypothesize a negative effect of the failure-related share of the rating variance on prices and sales. In this section, we empirically test these hypotheses using data on online ratings for digital cameras from Amazon.com.

4.1 Data

We obtained a data set from Amazon.com containing all single online reviews of 840 digital cameras (McAuley et al. 2015a, McAuley et al. 2015). The data were collected in July 2014 and contain all of the cameras' reviews accumulated by that time. Based on the online reviews, we computed the average ratings ($AVGRATE$), the number of reviews ($NUMREV$), the variance of ratings ($TOTALVAR$), the average length of the review texts ($AVGLEN$), and the average number of helpfulness votes a camera's reviews received ($TOTALHELP$) on a monthly basis (Table 2).

We obtained curve charts of sales ranks and prices over each camera's product lifetime from www.camelcamelcamel.com using a customized web crawler in November 2017. Using a self-developed software, we extracted all sales rank and price information from the charts and aggregated them on a monthly level. The top panels of Figure A1 and A2 in the appendix display a sample curve chart for the daily prices and sales ranks of a camera from www.camelcamelcamel.com. The bottom panel displays a plot of the monthly averages we extracted from the curve charts.

Table 2. Descriptive Statistics

Variable	N	Mean	Std. Dev.	Min	Max
SALES_RANK	44,240	12,335.12	11,087.5	0.5	135,683.1
FAILVARPERC	42,865	9.08	9.08	0	82.28
PRICE	44,240	304.25	304.25	.98	10016.39
AVG_RATING	43,877	4.03	0.53	1	5
NUM_REVIEWS	44,240	19.87	24.75	0	280
TOTALVAR	43,877	1.23	0.74	0	4
HELPFUL_VOTES	44,240	314.49	495.57	0	6493
AVG_LENGTH	43,877	1064.58	463.72	0	10389.5
Δ SALES_RANK	22,119	68.40	9082.83	–132,779	77567.21
Δ FAILVARPERC	20,904	0.91	24.67	–80	80.55
Δ PRICE	22,119	26.18	572.35	–9849.73	9896.49
Δ AVG_RATING	21,814	0.27	0.69	–4	3.6
Δ NUM_REVIEWS	22,119	0.29	26.84	–167	266
Δ TOTALVAR	21,814	0.02	0.9926	–3.37	4
Δ HELPFUL_VOTES	22,119	36.23	568.95	–3012	6268
Δ AVG_LENGTH	21,814	54.47	598.07	–9879.17	6522.25

Note: Prices are in $US and *AVG_LENGTH* is measured in the number of characters.

4.2 Identification of Product Failure Ratings

One crucial task in our study is to identify failure ratings in our large data set of online reviews. To this end, following prior literature (e.g., Tirunillai and Tellis 2014, Büschken and Allenby 2016, Debortoli et al. 2016), we employ probabilistic topic modeling based on LDA for our online review data set. LDA is a widely used unsupervised machine learning method that can identify topics in large collections of documents—in our case online reviews—with written text (Blei 2012, Debortoli et al. 2016). The essential idea behind LDA, according to Blei (2012), is that the authors compose documents D by first deciding about a discrete distribution of topics T to write about, and then they rely on words W from a discrete distribution of words that are typical for the chosen topic. Put differently, a document is defined by a probability distribution over a fixed set of topics, and each topic is defined by a probability distribution over a limited set of words (Debortoli et al. 2016). For each topic of the fixed set of topics, the LDA assigns a probability between 0 and 1 to each document (in this case: online review), indicating how likely it is that this particular document belongs to a certain topic.

The LDA approach has several advantages over alternative approaches that identify topics in written text. First, this approach can handle numerous documents in a very short time. Prior literature, such as in the field of marketing, has traditionally relied on manual coders to identify topics in online reviews (Sridhar and Srinivasan 2012). This approach is time consuming, costly, and difficult to replicate. Our approach circumvents these limitations. Second, it seems plausible that our underlying data suits the LDA assumption that there is a fixed set of topics underlying the documents. Accordingly, recent studies have highlighted the suitability of LDA to analyze online reviews (Debortoli et al. 2016). Even though one could argue that we could curate a training data set by hand and run supervised machine learning algorithms (e.g., support vectors machines) to detect failure ratings, LDA has the advantage that it also identifies various other topics in reviews that we leverage in the robustness checks of this study.

Before running the LDA using the web service minemytext.com, we applied standard measures of data preprocessing as suggested in the literature (Debortoli et al. 2016). In particular, we applied stemming to reduce the words to their stem, we used a total of 316 standard stop words (including *the*, *now*, *of*, and *and*, for example) to eliminate common and uninformative words, and we set the n-gram to 1. Because the LDA relies on a fixed number of topics that have to be determined by the researcher before running the analysis, the results obtained can depend on the number of topics chosen. As suggested in the literature (Debortoli et al. 2016), we tried several specifications with 15 to 80 topics. We evaluated the quality of the results by reading samples of the reviews marked as failure ratings and comparing the mean rating of the failure ratings with the remaining ones. The LDA that yielded the best results after visual inspection comprises 40 topics, which is within the range of 10 to 50 topics usually proposed in the literature (Debortoli et al. 2016). Table 3 provides examples that have been classified as failure ratings by our LDA approach as well as the respective rating given by the reviewer.

Within the LDA specification of 40 topics, exactly one topic captured the failure ratings, which we identified by reading the most frequently occurring words. Those words were camera (7.19%), len(se) (3.02%), problem (2.77%), repair (2.3%), replace (1.35%), fix (1.3%), and defect (0.51%). Because each review has a probability between 0 and 1 for each topic, we identified failure ratings as those reviews that had the highest probability by a 5% margin, among all possible topics, for the above-mentioned topic (failure).[2] As expected, the mean rating of the failure ratings (2.13) is substantially lower than the mean rating of the reviews excluding failure ratings (4.24). Figure 1 shows the distribution of failure and nonfailure ratings.

Table 3. Selected Reviews Classified as Failure Ratings by the LDA

Text	Rating
"Shutter button falls off and has a focus lock problem that should have resulted in a product recall. Sony then charges $120 for a repair."	1
"...bought this for when I moved out of the country for a year. The price was good for something that was supposed to tide me over while abroad. It worked, until it stopped working... The lens stopped protruding when the camera was turned on. I do not believe it was anything I did, as I did not drop it or toss it around. I kept it in a camera case. I know it must have been an uncommon defect or event, but I was very disappointed - and cameraless."	2
"I had high hopes for this camera when it arrived today, unfortunately for me the one that I ordered is crazy! The first few times I pressed the power button it would not turn on. After finally getting it to come on I tried taking pictures with it, during the first shot the camera cut off and I couldnt get it to come back on. The camera comes with 2AA duracell ultra batteries and the display showed that it was full when I got it to come back on. I wished I hadnt desired to save money since at the time there was another camera on sale for about twenty more bucks. I had found another of these when searching similar items that is about the same price but after checking targets (the seller) website for reviews I found one in which a reviewer had the exact same problem I am having and theirs ended up with the lens stuck out and useless after only one week, until I read that I had considered giving this one until my birthday at the beginning of the year to act right before returning it.[…]."	2

As depicted in Figure 1, the shape of the nonfailure rating distribution resembles the classical J-shape of online reviews (Hu et al. 2017) even though there are fewer 1-star ratings than expected,

[2] We can also impose more conservative requirements on a review's probability to qualify for a topic. For example, we can impose the restriction that a review only belongs to a certain topic, if the highest probability is 10% higher than the second highest probability, to avoid misclassifying reviews with extremely close values for the highest and second highest probability. Our baseline results reported in Table 5 remain qualitatively unchanged when we use different cutoffs (>0, 5, 10, 15, or 20%) for the review topic classification.

probably due to filtering out the failure ratings. The distribution of the failure ratings resembles a perfect *inverted* J-shape. Therefore, failure ratings are associated with substantially lower ratings than nonfailure ratings, probably due to the bad quality and the negative consumption experiences of the customers. However, some 4- and 5-star ratings might occur if product failures happen after a longer period of consumption that satisfy the customer.

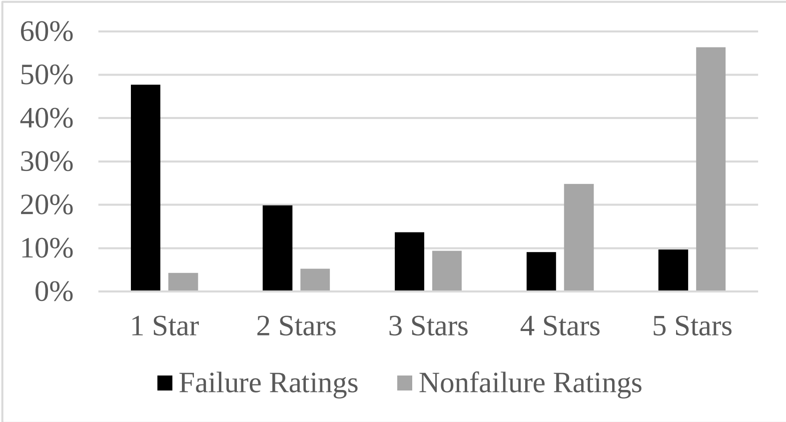

Figure 1. Distribution of Failure and Nonfailure Ratings

4.3 Validation of Identification

In a next step, we would like to validate the results of our topic modeling LDA approach. Our goal is to rule out concerns that our algorithm, for instance, misses capturing some failure ratings or mistakenly classifies ratings as failures that do not mention a product failure. Also, we would like to rule out that irony, sarcasm, or jargon causes our algorithm to misclassify a rating as a failure rating, even though the reviewer was not talking about a product failure. An effective way to do this is by human manual coding and calculating the interrater reliability between the human coding results and the LDA results. Here, we relied on manual human coding conducted independently by two research assistants. The assistants were asked to read a sample of reviews (without other information such as the rating or the brand) and determine whether the reviews mentioned product failures. Because the data contained too many reviews to be handled by humans, we took a random sample of our data comprising 300 reviews (following Lombard et al. 2002). Parametric and nonparametric tests of variables such as $SALES\_RANK$, AVG_RATING, and $PRICE$ did not show statistically significant differences between the entire data set and the random sample; thus, we concluded that the random sample was representative of the whole sample. The results of the agreement between the LDA and the human coders are depicted in Table 4.

Table 4. Interrater Agreement between LDA and Human Coders

	%-Agreement	N	Krippendorff's Alpha	Cohen's Kappa
LDA & C 1	97.99%	300	0.740	0.740
LDA & C 2	98.33%	300	0.806	0.806
LDA & C1 & C2	99.00%	300	0.808	0.875

At first glance, one can see that the agreement between the LDA and coder 1 (C1) and 2 (C2) is high (between 97.99% and 99%). This is reconfirmed by Krippendorff's Alpha and Cohen's Kappa, two conservative standard indices for evaluating interrater agreement (Lombard et al. 2002). Levels of 0.7 for K's Alpha and C's Kappa are sufficient to conclude interrater agreement, whereas levels above 0.8 indicate large interrater agreement (Lombard et al. 2002). Based on the results of interrater agreement between the human coders and the LDA, we concluded that our automatic topic modeling approach had reliably captured almost all failure ratings contained in our data set. This also reassured us regarding our choice of the number of topics (40) and the data preprocessing steps; these choices have obviously enabled a robust identification of failure ratings.

4.4 Main Variables

Finally, after the identification of failure ratings, we could compute the percentage share of failure ratings (*FAILVARPERC*) in the total online rating variance. As shown in equation (1), we calculated *FAILVARPERC* as follows. First, we calculated the taste-related variance (*TASTEVAR*). Then we divided the *TASTEVAR* by the total variance (*TOTALVAR*) to obtain the share of taste-related variance in the total variance. We computed *TASTEVAR* as the squared standard deviation from the mean of the nonfailure ratings (Γ, with $\gamma = 1,2, \dots \Gamma$) divided by the number of all ratings (N, with $n = 1,2, \dots N$), and we computed TOTALVAR as the variance of all ratings (Pindyck and Rubinfeld 2005). r represents a single rating, type t identifies nonfailure reviews, and μ represents the arithmetic mean of ratings. Because the remaining share of the variance had to be, by our definition, failure related, we subtracted the taste-related share of the online rating variance from 1 to obtain the failure-related share. We multiplied the result by 100 to obtain a percentage value between 0 and 100.

$$FAILVARPERC = 100 * \left(1 - \frac{TASTEVAR}{TOTALVAR}\right) \quad (1)$$

$$FAILVARPERC = 100 * \left(1 - \frac{\left(\frac{1}{N}\sum_{\gamma=1}^{\Gamma}(r_t - \mu_t)^2\right)}{\left(\frac{1}{N}\sum_{n=1}^{\mathrm{N}}(r - \mu)^2\right)}\right) \quad (2)$$

Our two dependent variables Y_{im}, are the difference in prices and in sales ranks (as a proxy of demand, in line with prior literature e.g., Chevalier and Mayzlin 2006, Garg and Telang 2013, Brynjolfsson et al. 2003) of a camera pair.[3]

4.5 Empirical Identification Strategy

Quite a number of existing studies have tried to establish a causal relationship between various metrics of online ratings (the average rating, the number of ratings, and to a lesser extent, the variance of ratings). All these studies, by and large, recognize the same endogeneity issue in that the causal effect of online ratings on economic outcomes can be biased by the unobservable quality of a product or service. The intuition is that if a product's quality is high this will simultaneously effect economic outcomes and online ratings in a positive way, obstructing the identification of the effect of online ratings on outcomes. Yet, studies using observational data have relied on a remarkable diverse set of strategies to mitigate endogeneity concerns. Some studies have relied on instrumental variable (IV) strategies (e.g., Jabr and Zheng 2014, Chintagunta et al. 2010), some studies have used regression discontinuity (RD) designs (e.g., Anderson and Magruder 2012, Luca 2016), other studies have used a simple fixed-effects (FE) specification (e.g., Einav 2007), and some studies acknowledge that they cannot mitigate endogeneity (e.g., Clemons et al. 2006).

Our study shares the endogeneity issue in that the identification of the causal impact of $FAILVARPERC$ on the sales and the price of a camera is obstructed by camera's unobservable objective quality. A camera's quality could determine the sales and the price of a camera and simultaneously it determines the probability of a camera to malfunction. A final set of existing studies has relied on differencing out confounding effects of unobservable product quality (e.g., Chevalier and Mayzlin 2006, Zhu and Zhang 2010, Sun 2012). Our identification approach is most closely related to these studies.

[3] We did not log-transform the variables because we did not find any significant skewness in the distribution.

Within our data set, we identified 420 pairs of camera models who differ from each other by mere time-constant arbitrary features such as color or the fact that the exact same model is sold alone or in a bundle with a tripod, for example.[4] Amazon.com curates separate product IDs, product pages, and separate reviews for these products. To isolate the causal effect of changes in $FAILVARPERC$ on prices and sales ranks, following the literature (e.g., Chevalier and Mayzlin 2006, Zhu and Zhang 2010), we subtracted all variables from each other, which differences out all monthly-constant unobservable variables ($\vartheta^{A/B}$, given that $\vartheta^{A} = \vartheta^{B}$) that might influence our dependent or independent variables (e.g., camera quality, brand popularity, or failure likelihood). This mimics a natural experiment in that the difference in the reported failure variance share for a camera pair is close to random. For example, the failure likelihood of the blue version of camera model X should be the same as for the green model. Thus, any difference in reported failures is nature's choice. Figure 2 displays an example of two identical camera models in our sample that merely differ in color. Technically, the differencing out strategy also removes unobservable seasonality bias for digital camera sales, if we assume that both cameras of a pair are equally affected by seasonality. To account for potential asymmetric seasonality effects, we can additionally implement monthly fixed effects.

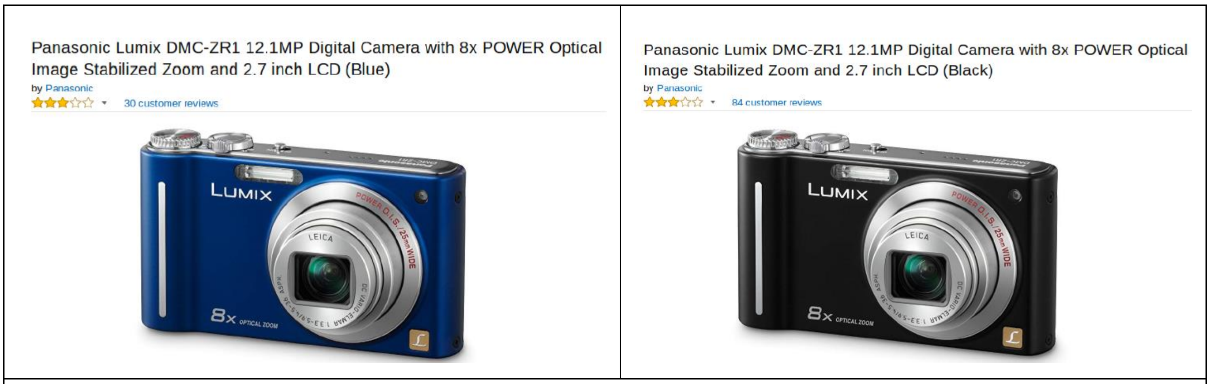

Figure 2. Example Camera Pair

Our identification strategy represents a natural way to leverage our data and circumvents possible pitfalls of other identification strategies (IV, RD, FE). Additionally, our approach differs to other differencing out strategies in a few meaningful ways. Chevalier and Mayzlin (2006) and Sun (2012) relied on a matched data set of identical books from two different online stores, which imposes the

[4] We manually match the pairs of cameras with replacement which results in 420 unique camera pairs.

assumption that both stores have relatively comparable customer populations and that online ratings are equally important and well-designed on both platforms. Zhu and Zhang (2010) use video games that are available for two different video game consoles which requires, for example, the assumption of comparable user populations, comparable game quality across consoles, and comparable review system design. The fact that we use data from the same websites alleviates many of these concerns. Yet, a disadvantage compared to the other approaches might be that our camera models are just *almost* exactly identical and not exactly identical (as in the case of Chegalier and Mayzlin (2006) or Sun (2012)).

4.6 Empirical Model

To investigate the effect between the failure-related variance on the price as well as the demand of a digital camera, we can estimate the following model (equation (2)) where the vectors X_{im} of camera pair i contain control variables (Table 2), θ_i captures camera-pair fixed effects, δ_m captures monthly fixed effects and ε_{im} is a random error term.

$$Y_{im}^{A} - Y_{im}^{B} = \beta_0 + \beta_1(FAILVARPERC_{im}^{A} - FAILVARPERC_{im}^{B}) + \beta_2(X_{im}^{A} - X_{im}^{B}) + (\vartheta^{A} - \vartheta^{B}) + \theta_i + \delta_m + \varepsilon_i \quad (3)$$

The key variable of interest in equation (2) is the variable $FAILVARPERC$, which indicates the share of the failure-related online rating variance of the total online rating variance. Therefore, this variable can take values between 0 and 100, and the coefficient β_1 captures the magnitude of the effect of the failure-related variance share on prices—respectively, sales.

4.7 Empirical Analysis

Table 5 displays the regression estimates for our regression model displayed in equation (3). Column (1) and (2) present the results of the model with $PRICE$ as the dependent variable, and column (3) and (4) present the results for the model with $SALES\_RANK$ as the dependent variable. In Column (1) and (3) we run a reduced form of our baseline model – only controlling for the total variance – and in models (2) and (4) we include all our control variables.

First, the results of column (2) show that the coefficient for $FAILVARPERC$ is negative and statistically significant. The magnitude of the coefficient (β_1: –2.35, s.e. 1.11) suggests that the share of failure-related variance negatively affects the price of a digital camera. Therefore, we find support for Hypothesis 1. It is important to note that we hypothesize the relationship of both variance components—the taste related and failure related—with the dependent variable. However, both relationships are measured by $FAILVARPERC$ because a 1% increase in the failure-related share of the variance automatically implies a 1% decrease in the taste-related share of the variance. In other words, a 1% increase in the taste-related variance share leads to an increase in price of \$2.35, on average. Also, unsurprisingly, the sign of the coefficient of the total variance is positive. The intuition behind this is that when keeping the failure-related share constant, the total absolute increase in the variance leads to an overproportional absolute increase in the taste-related variance compared to the failure related variance – as long as the failure-related share is below 50%, which increases the price.

Table 5. Baseline Results

Model	(1)	(2)	(3)	(4)
	PRICE	*PRICE*	*SALES_RANK*	*SALES_RANK*
FAILVARPERC	**–2.35****	**–2.37****	**49.18*****	**50.21*****
	(1.11)	(1.09)	(14.09)	(13.87)
PRICE				–0.18***
				(0.06)
TOTALVAR	78.47**	102.11**	166.13	–224.77
	(33.10)	(42.75)	(362.69)	(367.59)
NUMREV		0.19		–25.44**
		(0.40)		(11.63)
AVGRATE		41.76		–1,089.98*
		(34.64)		(579.07)
TOTALHELP		0.06*		–1.81
		(0.03)		(1.51)
AVGLEN		0.12***		–0.91
		(0.04)		(0.75)
Constant	16.88	5.14	–1,715.14**	–1,582.35**
	(11.36)	(12.70)	(781.64)	(785.94)
Camera-Pair Fixed Effects	✓	✓	✓	✓
Monthly Fixed Effects	✓	✓	✓	✓
Number of Groups	420	420	420	420
Observations	20,904	20,904	20,904	20,904
Adj. R²	0.00474	0.00608	0.10395	0.10777

Note: Robust standard errors are in parentheses. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

In particular, the magnitude of the coefficient of $FAILVARPERC$ is substantial. A 1% increase in the failure-related share of the online rating variance of a camera leads to a decrease in price by \$2.35. For instance, the absolute difference in the failure-related variance share between the 25th percentile and the 75th percentile of $FAILVARPERC$ is 7%. Therefore, the price of a camera from the 75th percentile of $FAILVARPERC$ is around \$16.45 higher than the price of a camera from the 25th percentile of

$FAILVARPERC$. Reassuringly, the signs of the coefficients of $PRICE$, $NUMREV$, and $AVGRATE$ are all in line with economic intuition and prior literature (e.g., Anderson and Magruder 2012, Chevalier and Mayzlin 2006).

Second, the results of column (4) show that the coefficient for $FAILVARPERC$ is positive and statistically significant. Thus, the sales rank for a digital camera increases – i.e., the demand decreases – when the share of the failure-related variance increases (β_1: –50.21, s.e. 13.87). Therefore, we also find support for Hypothesis 2. Again, analogous to this, the taste-related variance share has a negative effect on the sales rank. Also in this case, the magnitude of the coefficient is pronounced. In particular, a 1% increase in the failure-related share of the online rating variance of a camera leads to an increase in the sales rank by 50 ranks on average. For instance, the sales rank of a camera from the 75th percentile of $FAILVARPERC$ is around 350 higher than the sales rank of a camera from the 25th percentile of $FAILVARPERC$.

The sign of the coefficient of the total variance is negative, as expected; keeping the failure share in the variance constant, an increase in the total variance is associated with more sales. As explained earlier, this is because, when keeping the failure-related share constant, the total absolute increase in the variance leads to an overproportional absolute increase in the taste-related variance compared to the failure related variance – as long as the failure-related share is below 50%, which decreases the sales rank.

4.8 Robustness Checks

We conducted several robustness checks to ensure validity of our results.

First, one concern is related to the probability cutoffs we used to classify our online reviews into topics. In our baseline model, we classify an online rating as a failure rating if the review text has the highest probability for the failure topic by a 5% margin. Even though we validated our failure rating identification with human coders, one might argue that there might be a second candidate topic with a similarly high probability for this topic. In other words, there might be a second topic besides product failure contained in the review text. To mitigate these concerns, we used different probability margins (>0, 5, 10, 15, or 20%) for ratings to classify as failure ratings. Our results remain qualitatively unchanged when we run our baseline regression again with the newly-computed $FAILVARPERC$ based on these margins.

Table 6. Results Without Quality Reviews

Model	(1)	(2)	(3)	(4)
	PRICE	*PRICE*	*SALES_RANK*	*SALES_RANK*
FAILVARPERC	**–2.98***	**–3.10***	**39.13*****	**41.55*****
	(1.61)	(1.61)	(14.05)	(13.78)
PRICE				–0.19***
				(0.06)
TOTALVAR	89.49**	107.35**	133.09193	–39.62
	(44.27)	(49.18)	(355.79)	(341.58)
NUMREV		0.55		–32.51**
		(0.50)		(14.44)
AVGRATE		11.77		–291.78
		(28.63)		(526.90)
TOTALHELP		0.06		–2.97
		(0.04)		(1.82)
AVGLEN		0.15***		–0.80
		(0.05)		(0.64)
Constant	24.71**	13.71	–1,818.85611**	–1,722.78**
	(11.45)	(11.97)	(798.20)	(801.39)
Camera-Pair Fixed Effects	✓	✓	✓	✓
Monthly Fixed Effects	✓	✓	✓	✓
Number of Groups	412	412	412	412
Observations	20,340	20,340	20,340	20,340
Adj. R^2	0.00541	0.00733	0.10705	0.11073

Note: Robust standard errors are in parentheses. $^{*}p < 0.1$; $^{**} p < 0.05$; $^{***} p < 0.01$.

Second, one might be concerned about our implicit assumption that every review that was not a failure review was automatically said to increase the taste-related variance. Even though this classification is theoretically and logically valid, this may not guarantee that nonfailure reviews really contain taste-related information. While, ultimately, classifying the information contained in the nonfailure reviews is an intricate task, we aim at alleviating these concerns by dropping reviews whose topics are about vertical (quality-related) rather than horizontal (taste-related) features. The key idea is that every reviewer will appreciate higher quality. Thus, the role of taste should be negligible for reviews mentioning the quality aspects of a camera. Consequently, we dropped all the reviews from topics that had the word "quality" among the 7 most frequently occurring words. We did not drop reviews categorized as failure reviews. We re-estimated our baseline model after removing the "quality" reviews, but our results remained qualitatively unchanged (Table 6). This gives support to our classification of review in just two categories: taste- and failure-related reviews.

5 Discussion

To a certain extent, our results provide tentative explanations to reconcile conflicting findings of past studies in follow-up studies.

First, and most importantly, one potential explanation for the inconclusive findings of prior studies could be the fact that the analyses conducted in the respective studies did not differentiate between a failure-related and a taste-related share in the variance. As outlined in section 2, prior literature has focused on products such as craft beer and movies and services such as hotel stays, for instance. Both of them have two things in common: they comprise a wide range of experience attributes (hotels: friendliness of the staff, taste of the food, comfort of the bed; cameras: product failure, ease of use of the software, usability, feel of the product). Also, both of them can fail. But we recognize that the potential for service failure is likely larger than the potential for product failures with craft beer and movies, as documented in the literature (Webster and Sundaram 1998, Smith and Bolton 1998, Hess Jr. et al. 2003), especially for hotels (Proserpio and Zervas 2017). Consequently, product failure can be leveraged to decompose the total variance in these cases. Although it is impossible to judge ex-post if the different sources of the variance that have been neglected by prior studies constitute the driving force behind the conflicting results, this is one candidate explanation that should be explored in future research. Moreover, the fact that studies investigating the product sales of craft beer and movie tickets have not found a negative relationship between the online rating variance and sales might be due to a limited potential for these products to fail. Finally, a logical next step would be to re-evaluate the results of the studies focusing on electronics (Kostyra et al. 2016, Langan et al. 2017, Minnema et al. 2016, Wang et al. 2015) using our approach of distinguishing between taste- and failure-related variance.

Second, our findings are in line with Clemons et al. (2006) and Bao and Chang (2014), although we explore more differentiated facets of the rating variance, whereas they simply utilize the total variance. We also find that the overall variance is positively associated with sales. Even though beer and books represent experience goods, one can hardly imagine them to “fail.”

Third, our results deliver empirical support for prior experimental (He and Bond 2015) and theoretical studies (Zimmermann et al. 2017). He and Bond’s (2015) experimental study investigated the effect of the variance for taste-dissimilar (paintings) and taste-similar (niche computer games) products on purchase intention. They found that a high variance is beneficial for the purchase intention for taste-dissimilar products. In other words, for products that some potential customers like and some others dislike because they have different tastes, the variance increases the purchase intention (He and Bond 2015). For taste-similar products, most of the reviewers should have similar tastes (e.g., players of a certain type of video games). If players observe a high variance of online ratings, their purchase intention decreases because they associate this variance with low product quality. Our results also lend empirical support to this study. In our case, the taste-related variance is the extent to which a digital camera is a taste-dissimilar product. By the same token, the failure-related variance represents the extent to which a camera is taste similar; all customers agree that they dislike product failure.

6 Conclusion

Surprisingly few studies have investigated the relationship between the online rating variance and sales, and their results have been inconclusive. This lack of knowledge is a handicap for researchers, managers, and customers alike. Despite some theoretical accounts, little empirical consensus has been reached in the field of user-generated social media content and, in particular, online ratings. Thus, customer decision-making under customer disagreement has remained, by and large, a black box. Second, managers need to understand the implications of the online rating variance to adapt the sales strategies of their products and services. Without knowledge of the effects of the variance on strategic variables such as price and demand, managers cannot be sure if they should change their prices or alter their demand forecast. Third, although customers might intuitively know how to make sense of the online rating variance, interpreting the online rating variance in combination with the average rating has remained difficult (as for instance in Sun (2012)).

To close this knowledge gap, our study focuses on empirically breaking down the effects of two distinct components of the online rating – the taste-related share and the failure-related share – on prices and sales. We study this effect in the context of a consumer good— digital cameras—that can actually fail with a relatively large share of experience attributes. We build upon a theoretical model by Zimmermann et al. (2017) to delineate hypotheses that we later test on our data set of online ratings for digital cameras from Amazon.com. Our empirical results highlight that an increase in the product failure-related share in the online rating variance has a negative impact on prices and sales. The explanation for this is that all customers dislike the failure of a product, which lowers sales and the prices a retailer can charge. In contrast to that, an increase in the taste-related share of the online rating variance positively affects prices and sales. This type of customer disagreement helps customers find a product that they really like; when past customers give negative ratings about the difficult usability of a camera, advanced users might prefer this camera over another, simpler camera. Consequently, the retailer can charge higher prices and enjoy a higher demand.

Our results have substantial implications for research and practice. To the best of our knowledge, we are the first to empirically decompose the online rating variance of a product into a failure- and a taste-related component to study the ambiguous relationship of the variance with economic outcomes. Thus, our study helps scholars understand the economic implication of the rating variance. With regard to theory, our results lend empirical support to the model of Zimmermann et al. (2017). In addition, our study provides field-empirical support to the laboratory findings of He and Bond (2015).

Retailers can benefit from our results to better sell their products. They can adjust their price and demand forecasts based on the observed shares of taste-related and failure-related online rating variance of their product. Moreover, manufacturers can learn about the failures of their own products and improve their design based on the failure ratings.

Customers are able to better incorporate the online rating variance into their purchase decisions. For instance, customers might first inspect negative ratings to check whether they are based on product failure or based on taste-related aspects. Review systems should assist this process by facilitating the identification of failure ratings to help customers find the products they like best.

Naturally, this study also comes with limitations that present avenues for future research. Because our study focuses on consumer goods, a natural extension would be to conduct a study breaking down the online rating variance of services by identifying when services fail (Hess Jr. et al. 2003, Proserpio and Zervas 2017). This might help to explain the conflicting findings by studies on the online rating variance. Also, digital cameras represent a consumer good, and it is unclear how our results play out for digital goods such as apps or video games. Digital goods can also fail–due to bugs or crashes–and future research could study whether taste differences are also beneficial to prices and sales of these goods. Our empirical identification strategy of the causal relationship between the failure-related variance share and prices and sales is based on the assumption, that unobservable differences between two cameras of a matched pair are constant over a month. That is, if a blue model generally enjoys a higher demand than a black model, this does not bias our estimation as long as these differences are constant within a month and uncorrelated with *changes* the customers' willingness to pay. While we have no reason to believe that the customers' willingness to pay should exhibit any dynamics systematically related to our products during the course of a month, this could be investigated in a follow-up laboratory experiment.

References

Allen, R. (2016). Incredible New Stats Show Just How Important Reviews and Ratings Are for Ecommerce Sites, URL: https://www.smartinsights.com/ecommerce/customer-reviews-social-proof/incredible-new-stats-show-just-crucial-reviews-ecommerce-sites/ (visited 04/13/2018).

Anderson, M., and J. Magruder (2012). "Learning from the Crowd: Regression Discontinuity Estimates of the Effects of an Online Review Database," *The Economic Journal* 122 (563), 957–989.

Archak, N., Ghose, A. and P. G. Ipeirotis (2011). "Deriving the Pricing Power of Product Features by Mining Consumer Reviews," *Management Science* 57 (8), 1485–1509.

Bao, T., and T.-l. S. Chang (2014). "Finding Disseminators via Electronic Word of Mouth Message for Effective Marketing Communications," *Decision Support Systems* (67), 21–29.

Blei D. M. (2012). "Probabilistic Topic Models," Communications of the ACM 55 (4), 77–84.

Brynjolfsson, E., Hu, Y., and M. D. Smith (2003). "Consumer Surplus in the Digital Economy: Estimating the Value of Increased Product Variety at Online Booksellers." *Management Science*, 49(11), 1580-1596.

Büschken, J., and G. M. Allenby (2016). „Sentence-based Text Analysis for Customer Reviews." *Marketing Science*, 35(6), 953-975.

Cabral, L. and A. Hortaçsu (2010). "The Dynamics of Seller Reputation: Evidence from eBay," *The Journal of Industrial Economics* 58 (1), 54–78.

Chen, Y. and J. Xie (2008). "Online Consumer Review: Word-of-Mouth as a New Element of Marketing Communication Mix," *Management Science* 54 (3), 477–491.

Chevalier, J. A. and D. Mayzlin (2006). "The Effect of Word of Mouth on Sales: Online Book Reviews," *Journal of Marketing Research* 43 (3), 345–354.

Choi, S. Y., Stahl, D. O. and A. B. Whinston (1997). *The Economics of Electronic Commerce*. Indianapolis, IN: Macmillan Technical Publishing.

Chintagunta, P. K. Gopinath, S., and S. Venkataraman (2010). "The Effects of Online User Reviews on Movie Box Office Performance: Accounting for Sequential Rollout and Aggregation Across Local Markets," *Marketing Science* 29 (5), 944–957.

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

Debortoli, S., Junglas, I., Müller, O. and J. Vom Brocke (2016). "Text Mining for Information Systems Researchers: An Annotated Topic Modeling Tutorial," *Communications of the Association for Information Systems* 39 (7), 110—135.

Dellarocas C. (2003). "The Digitization of Word of Mouth: Promise and Challenges of Online Feedback Mechanisms," *Management Science* 49 (10), 1407–1424.

Drewnicki N. (2013). Survey: 90% Say Positive Reviews Impact Purchase Decisions, URL: https://www.reviewpro.com/blog/survey-zendesk-mashable-dimensional-research-90-say-positive-reviews-impact-purchase-decisions/ (visited 11/24/2017).

Duan, W., Gu, B. and A. B. Whinston (2008). "Do Online Reviews Matter?—An Empirical Investigation of Panel Data," *Decision Support Systems* 45 (4), 1007–1016.

Einav, L. (2007). Seasonality in the US Motion Picture Industry. *The Rand Journal of Economics*, 38(1), 127-145.

Feng, J., Li, X., and X. Zhang (2016). Online Product Reviews-Triggered Dynamic Pricing: Theory and Evidence. *Information Systems Research* (Forthcoming). URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2510470 (visited 09/14/2018).

Garg, R., and R. Telang (2013). Inferring App Demand from Publicly Available Data. *Management Information Systems Quarterly* 37(4), 1253-1264.

Gutt, D. (2018). "Sorting Out the Lemons–Identifying Product Failures in Online Reviews and their Relationship with Sales." Proceedings of the Multikonferenz Wirtschaftsinformatik (MKWI), Lüneburg, Germany, pp. 1649–1655.

He, S. X. and S. D. Bond (2015). "Why Is the Crowd Divided?: Attribution for Dispersion in Online Word of Mouth," *Journal of Consumer Research* 41 (6), 1509–1527.

Hess Jr., R. L., Ganesan, S. and N. M. Klein (2003). "Service Failure and Recovery: The Impact of Relationship Factors on Customer Satisfaction," *Journal of the Academy of Marketing Science* 31 (2), 127–145.

Hong, Y., Chen, P.-Y. and L. M. Hitt (2012). "Measuring Product Type with Dynamics of Online Product Review Variance," In: *Proceedings of the 33rd International Conference on Information Systems, Orlando, FL, USA*, pp. 2034–2051.

Hu, N., Pavlou, P. A. and J. Zhang (2017). "On Self-Selection in Online Product Reviews," *Management Information Systems Quarterly* 41(2), 449–471.

Jabr, W., and Z. E. Zheng (2014).“ Know Yourself and Know Your Enemy: An Analysis of Firm Recommendations and Consumer Reviews in a Competitive Environment.” *Management Information Systems Quarterly* 38(3), 635-654.

Joshi, A. and H. Mao (2012). “Adapting to Succeed?: Leveraging the Brand Equity of Best Sellers to Succeed at the Box Office,” *Journal of the Academy of Marketing Science* 40 (4), 558–571.

Koh, N. S., Hu, N. and E. K. Clemons (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 (5), 374–385.

Kostyra, D. S., Reiner, J., Natter, M., and D. Klapper (2016). “Decomposing the Effects of Online Customer Reviews on Brand, Price, and Product Attributes,” *International Journal of Research in Marketing* 33 (1), 11–26.

Kwark, Y., Chen, J. and S. Raghunathan (2014). “Online Product Reviews: Implications for Retailers and Competing Manufacturers,” *Information Systems Research* 25 (1), 93–110.

Langan, R., Besharat, A., and S. Varki (2017). “The Effect of Review Valence and Variance on Product Evaluations: An Examination of Intrinsic and Extrinsic Cues,” *International Journal of Research in Marketing* 34 (2), 414–429.

Lewis, G. and G. Zervas (2016). “The Welfare Impact of Consumer Reviews: A Case Study of the Hotel Industry,” Working Paper.
URL: http://faculty.chicagobooth.edu/workshops/marketing/pdf/tawelfare.pdf (visited 11/27/2017).

Li, X. and L. M. Hitt (2008). “Self-Selection and Information Role of Online Product Reviews,” *Information Systems Research* 19 (4), 456–474.

Lombard, M., Snyder-Duch, J. and C. C. Bracken (2002). “Content Analysis in Mass Communication: Assessment and Reporting of Intercoder Reliability,” *Human Communication Research 28* (4), 587–604.

Lu, Q., Qiang, Y. and R. Law (2014). “Moderating Effects of Product Heterogeneity Between Online Word-of-Mouth and Hotel Sales,” *Journal of Electronic Ecommerce Research* 15 (1), 1–12.

Luca M. (2016). “Reviews, Reputation, and Revenue: The Case of Yelp.Com,” URL: https://papers.ssrn.com/sol3/Papers.cfm?abstract_id=1928601 (visited 11/27/2017).

McAuley, J., Pandey, R. and J. Leskovec (2015a). “Inferring Networks of Substitutable and Complementary Products,” Proceedings of the 21th In: *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794.

McAuley, J., Targett, C., Shi, Q. and A. van den Hengel (2015). “Image-Based Recommendations on Styles and Substitutes,” In: *Proceedings of the 38th ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 43–52.

Meiseberg B. (2016). “The Effectiveness of E-tailers’ Communication Practices in Stimulating Sales of Niche Versus Popular Products,” *Journal of Retailing* 92 (3), 319–332.

Minnema, A., Bijmolt, T. H.A., Gensler, S., and T. Wiesel (2016). “To Keep or Not to Keep: Effects of Online Customer Reviews on Product Returns,” *Journal of Retailing* 92 (3), 253–267.

Moon, S., Bergey, P. K. and D. Iacobucci (2010). “Dynamic Effects Among Movie Ratings, Movie Revenues, and Viewer Satisfaction,” *Journal of Marketing* 74 (1), 108–121.

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

Netzer, O., Feldman, R., Goldenberg, J. and M. Fresko (2012). “Mine Your Own Business: Market-Structure Surveillance Through Text Mining,” *Marketing Science* 31 (3), 521–543.

Neumann, J. and D. Gutt (2017). “A Homeowner’s Guide to Airbnb: Theory and Empirical Evidence for Optimal Pricing Conditional on Online Ratings,” In: *Proceedings of the 25th European Conference on Information Systems (ECIS), Guimaraes, Portugal*, pp. 997–1010.

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

Pindyck, R. and D. Rubinfeld (2005). Microeconomics, Upper Sadde River, NJ: Prentice Hall.

Proserpio, D. and G. Zervas (2017). “Online Reputation Management: Estimating the Impact of Management Responses on Consumer Reviews,” *Marketing Science* 36 (5), 645–665.

Raguseo, E., Neirotti, P., and E. Paolucci (2017). “How Small Hotels can Drive Value their Way in Infomediation. The Case of ‘Italian Hotels vs. OTAs and TripAdvisor’,” *Information & Management* 54 (6), 745–756.

Schlosser, A. E. (2011). “Can including pros and cons increase the helpfulness and persuasiveness of online reviews? The interactive effects of ratings and arguments.” *Journal of Consumer Psychology* 21(3), 226-239.

Sen, S. and D. Lerman (2007). “Why Are You Telling Me This? An Examination into Negative Consumer Reviews on the Web,” *Journal of Interactive Marketing* 21 (4), 76–94.

Shapiro C. (1983). “Optimal Pricing of Experience Goods,” *The Bell Journal of Economics* 14 (2), 497–507.

Smith, A. K. and R. N. Bolton (1998). “An Experimental Investigation of Customer Reactions to Service Failure and Recovery Encounters: Paradox or Peril?” *Journal of Service Research*, 1 (1), 65-81.

Sridhar, S. and R. Srinivasan (2012). “Social Influence Effects in Online Product Ratings,” *Journal of Marketing* 76 (5), 70–88.

Sun M. (2012). “How Does the Variance of Product Ratings Matter?” *Management Science* 58 (4), 696–707.

Teubner, T., Hawlitschek, F. and D. Dann (2017). “Price Determinants on Airbnb: How Reputation Pays Off in the Sharing Economy,” *Journal of Self-Governance and Management Economics* 5 (4), p. 53—80.

Tirunillai, S., and G. J. Tellis, G. J. (2014). "Mining Marketing Meaning from Online Chatter: Strategic Brand Analysis of Big Data using Latent Dirichlet Allocation." *Journal of Marketing Research*, 51(4), 463-479.

Wang, F., Liu, X., and Fang, E. 2015. "User Reviews Variance, Critic Reviews Variance, and Product Sales: An Exploration of Customer Breadth and Depth Effects," *Journal of Retailing* 91 (3), 372–389.

Webster, C. and D. S. Sundaram (1998). Service Consumption Criticality in Failure Recovery. *Journal of Business Research* 41 (2), 153—159.

Wu, J., Wu, Y., Sun, J., and Z. Yang (2013). "User Reviews and Uncertainty Assessment: A Two Stage Model of Consumers' Willingness-to-Pay in Online Markets," *Decision Support Systems* 55 (1), 175–185.

Ye, Q., Law, R. and B. Gu (2009). "The Impact of Online User Reviews on Hotel Room Sales," *International Journal of Hospitality Management* 28 (1), 180–182.

Ye, Q., Law, R., Gu, B. and W. Chen (2011). "The Influence of User-Generated Content on Traveler Behavior: An Empirical Investigation on the Effects of E-Word-of-Mouth to Hotel Online Bookings," *Computers in Human Behavior* 2 (2), 634–639.

Zimmermann, S., Herrmann, P., Kundisch, D. and B. Nault (2017). "Decomposing the Variance of Consumer Ratings and the Impact on Price and Demand," *Information Systems Research* (Articles in Advance).

Appendix

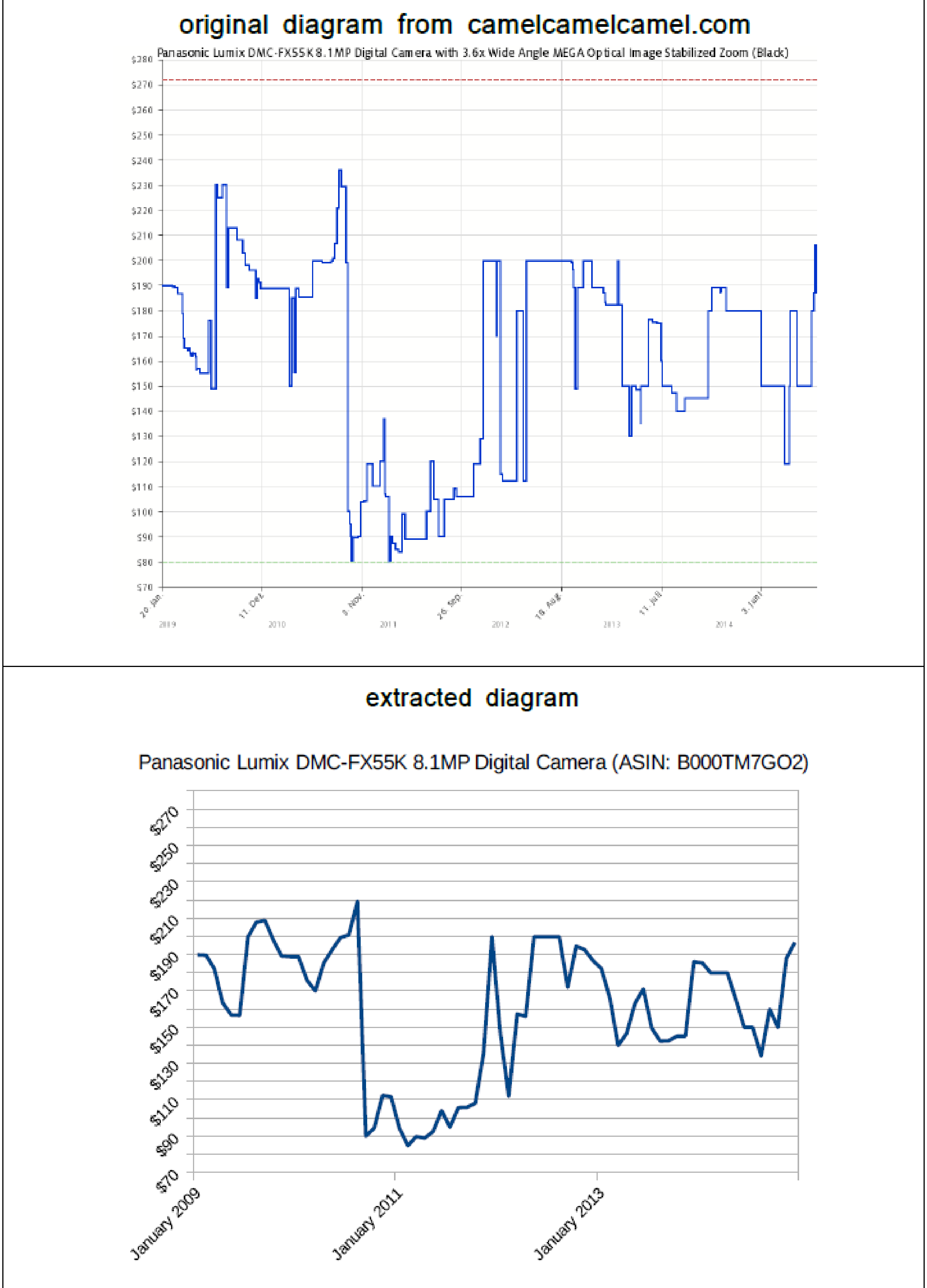

Figure A1. Crawled Curve Chart of Price Data (Top Panel) and Plotted Monthly Averages (Bottom) original salesrank diagram from camelcamelcamel.com

extracted diagram

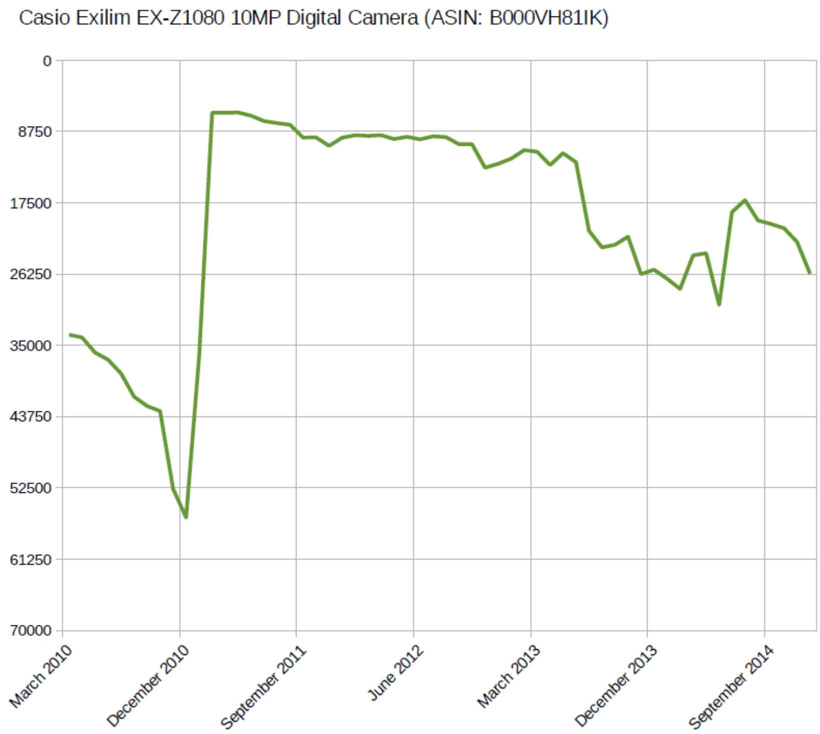

Figure A2. Crawled Curve Chart of Sales Rank Data (Top Panel) and Plotted Monthly Averages (Bottom)