ERASMUS UNIVERSITY ROTTERDAM
Erasmus School of Economics

MSc. Economics and Business
Specialization Data Science & Marketing Analytics

Owning the conversation on your product and brand: a study on online product review valence across different types of online marketplaces

Author: Tom van den Broek 444118
Supervisor: Dr. M. van Crombrugge
Co-reader: Dr. A.J. Koning

July 15, 2021

Aknowledgements

I would like to thank everyone who has helped me over the past five years in completing my studies. During my time at the Erasmus School of Economics, I have learned how to think and work like an economist, and I can truly say that my time here has transformed me.

It is therefore fitting to finish my studies with this thesis that has allowed me to apply the lessons I have learned to a topic that really arouses my interest. I thoroughly enjoyed writing it, and I hope that it is enjoyable to its readers and that its results can spark others to look more into this topic.

A special thanks goes out to my thesis supervisor Dr. Michiel van Crombrugge, who encouraged me from the start to research a topic that few authors in the field have looked at, and whose constructive feedback guided me through this intense process. Finally, I would like to thank my family and friend for supporting me throughout my academic career.

Abstract

In the future e-commerce will only expand its already dominant position in the retail landscape. As such, manufacturers will have to formulate and rethink their digital sales strategies. A manufacturer can take matters into its own hands and focus on its own online direct channel, but e-commerce giants like Amazon present a great opportunity to generate more traffic and sales. However, this also implies a loss of ownership over the customer relationship. Among other things, brand managers no longer influence how customers perceive their products and brand in these marketplaces. This study looks at one aspect of this experience: the online product reviews. We study how their valence, which is known to help or harm a manufacturer's brand and sales, differs over marketplaces that are owned by the manufacturer or not. Using sentiment analysis we find that reviews are on average more positive in terms of sentiment in the channel owned by the manufacturer, but only find evidence that the difference in price between the two can affect this difference in sentiment. With our findings we contribute to the academic field studying manufacturer encroachment, by adding a new motive why a manufacturer would want to focus on its direct online channel. Moreover, we add value for brand managers who are rethinking their (online) sales channel composition, by giving empiric proof for a reason to focus on their own channels.

Keywords: Manufacturer encroachment, online sales channels, online product reviews, sentiment analysis

Contents

1 Introduction

In 2020 the COVID-19 pandemic changed the way consumers purchase their products substantially. Forced closures of physical retail stores meant that e-commerce firms such as Amazon and Alibaba experienced enormous growth in sales (OECD, 2020; Sahli, 2020). Though this spike was temporary, many analysts expect this move toward online sales to be permanent as consumers get used to the convenience of this way of shopping (Sides & Skelly, 2021). This trend obviously has many implications for traditional brick-and-mortar retailers, but it also forces manufacturers to rethink their strategy toward their sales channels. Among other things, they need to have a clear digital sales strategy (PwC, 2020). It would seem obvious to double down on sales on e-commerce platforms such as Amazon, but not all firms follow this path. Already before the pandemic, sportswear manufacturer Nike decided to pull its products from Amazon (Novy-Williams & Soper, 2019). The company indicated that it wanted to focus on its own online selling platform, wanting to provide customers with a more direct and personal relationship with the brand. This move of Nike taking its online sales in its own hands may be a natural consequence of the shift of sales towards e-commerce: as consumers get their products delivered to their doorstep, there is less need for 'boots on the ground' in terms of retail channels. This may allow manufacturers to put more emphasis on their own, direct channels. In the case of Nike, one of the reasons for its move was the fact that it wanted to directly control the way consumers engage with its brand (Novy-Williams & Soper, 2019).

The way consumers experience and interact with a brand is really important for a firm as its brand is its most valuable asset (Aaker, 2009). Marketers can positively influence their brand image through advertisements, but with the rise of the internet they no longer control all communications on their brands (Yu, Liu, Lee, & Soutar, 2018). Take for instance social media: firms can communicate with their customers here, but it has also been shown that user-generated communications significantly influence the firm's brand image as well (Bruhn, Schoenmueller, & Schäfer, 2012). This paper focuses on customer reviews, a form of such online communications about brands and their products taking place in online sales channels. This content can help or harm a brand, as it has been shown to affect product sales and brand image (Lin & Xu, 2017). More specifically we focus on the valence of these reviews, as many studies have found that this affects sales and brand image substantially (Bambauer-Sachse & Mangold, 2011; Chevalier & Mayzlin, 2006; Cui, Lui, & Guo, 2012).

We aim to establish how a company has its brand experience affected when its products are sold in third party online marketplaces that it has no control over. By leveraging text mining techniques to analyze the valence in online product reviews, we aim to empirically show how these reviews might make the way consumers experience a brand or product different between a firm's own direct online sales channel and a third party e-commerce platform like Amazon. In sum, we answer the following research question:

How does online product review valence differ between direct online sales channels and third party online sales channels?

This research question can be divided in the following sub-questions, which we answer to arrive at our conclusions:

- How do online product review sentiments differ between a direct and third party online sales channel?

- How do online product review ratings differ between a direct and third party online sales channel?
- How is this difference moderated by product level factors?
- How is this difference moderated by review level factors?

We find that, consistent with our conceptual framework, reviews in the marketplace owned by the brand are on average more positive in terms of sentiment compared to third party marketplaces. However, this only holds when looking at valence through a review's sentiment, not its star rating. For the moderating effects we expected we find little evidence, only that the difference in price between the two can affect the difference in sentiment.

The findings presented in this paper are relevant for large firms that place great value on their brand. More specifically, they function as empiric evidence that can help these firms that have to make important decisions regarding their future digital strategies (PwC, 2020). The choice to partner up with an e-commerce giant like Amazon presents a trade-off: it presents a great opportunity to generate sales and traffic, but it also implies a loss in control over the way consumers interact with a brand and its products, for instance through online product reviews (Novy-Williams & Soper, 2019; Zimmerman, 2020). For smaller companies this extra traffic may be worth it, but even among these parties much attention is given to managing online product reviews (Chen, 2019). For larger brands the damage due to the loss of control may be too much, as online product reviews can decrease the role of the brand image they heavily invest in (Aaker, 2009; Kostyra, Reiner, Natter, & Klapper, 2016). It is for these brands' managers that we provide empiric evidence to support their decision making: by showing whether or not and potentially why the review valence differs between their own marketplace and Amazon, our findings can tell them if these reviews are a relevant factor to consider in their digital sales strategy choices.

This paper continues with a literature section, where we show what has already been done in the field and how we contribute to it. We then present our theoretical framework, where we review theory to identify the variables relevant for this study and the corresponding hypotheses. After this we explain how we obtained the data for this analysis followed by the methodology used to perform it. We then show the results and end the paper with a conclusion and a discussion of our findings.

2 Contributions

In this paper we investigate a manufacturers' motive to shift more of its sales to its online direct sales channel. This motive is the fact that the manufacturer has more control over the user experience on its own platform. There already is an academic field on a manufacturer encroachment: a manufacturer taking sales to its direct channel. However, this field has no works yet with a focus on this specific reason for it. In a seminal piece, Chiang, Chhajed, and Hess (2003) mainly focus on the consequences on prices when a direct channel is set up next to a retail channel, finding that this direct channel leads to more optimal pricing which benefits both the manufacturer and retailer. Arya, Mittendorf, and Sappington (2007) take a different viewpoint by looking at the effects of the introduction of a direct channel from the retailer's perspective, and find the same conclusion. Tsay and Agrawal (2004) take a more scenario-based approach and add to the above conclusions that it depends heavily on the circumstances. For instance, when a manufacturer has

strong marketing capabilities, it should consider setting up a direct channel only for marketing purposes but diverting the final sales to a reseller. Finally, Cai (2010) considers the role having a direct channel could have in negotiating contracts with retailers, as the threat of a direct channel replacing the retailer's sales could benefit a manufacturer in negotiating a higher revenue share.

An addition to the field by Kumar and Ruan (2006) already relates more to the motive that we study. They consider rival products as well, and find that a direct channel is less optimal when a product requires the retailer to support customers in the shopping experience. In this case, a manufacturer must enhance retail support to beat its competitors in stores. This finding can be traced back to the reason for a manufacturer to prefer its own online channel over a third-party platform: as there are no 'boots on the ground' in either one and there is hardly any support element in play, the manufacturer might as well take the online customer experience in its own hands. This paper thus already hints at the motive that we investigate in this paper, but not in an online shopping environment.

Very recent work by W. Yang, Zhang, and Yan (2021) in this field has actually considered how online product reviews may affect this choice of sales channel. However, they still only study a dual channel consisting of a direct online channel and a traditional brick-and-mortar retail channel. They study online product reviews as an aspect of the direct channel to be considered by the manufacturer, and what role this choice plays in the relationship between manufacturer and retailer. Among other things, they find that it is not necessarily wise for the manufacturer to have these reviews present in its direct online channel, unless the content of those reviews is beneficial for the manufacturer. Even though the focus of this work is different, it supports the idea that online product reviews and their valence are actually important to consider in the choice of sales channels. However, no work to date has investigated online product reviews as a factor in the choice between an own direct online channel and a third-party *online* sales channel. This is where we want to contribute to the literature, by providing empiric proof that being able to control the online product reviews written about your products is indeed a motive to prefer online selling through your direct channels.

3 Theoretical Framework

3.1 Online product reviews

This paper focuses on product customer reviews that accompany online product pages. These reviews are part of a larger stream of User Generated Content (UGC) or electronic Word Of Mouth (eWOM) that also include social media conversations and forum discussions (Homburg, Ehm, & Artz, 2015; Liu, Burns, & Hou, 2017). They are present in nearly all online marketplaces today as they can help customers in choosing products they cannot touch or smell as they would in offline stores (Park, Lee, & Han, 2007). As such these reviews can reduce consumers' choice risk (Kostyra et al., 2016). Park et al. (2007) present online product reviews as having a dual role in the customer journey. First, they have an informational role, as they function as a channel for product information that complements the information given by the seller. Second, they have a recommender role where previous users can recommend a product or not. In both these roles they derive their power from their trustworthiness, as the reviews are written by previous customers voicing their honest opinions. Given their importance in the customer decision making process, many works exist on this topic. By far the largest stream of work focuses on the consequences that these reviews have on firms. We

summarize these below.

3.2 Impact of online product reviews on firms

The most studied influence of online product reviews is their influence on customer purchase intention. Seminal work by Chevalier and Mayzlin (2006) studies online book reviews and finds that when a book's reviews improve its sales also increase, and that the effect of an overwhelmingly negative review is larger than that of a positive review. This is an intuitive but very significant finding, and subsequent works have looked at this effect of a review's valence on purchase intention in more detail. While many confirm this basic positive relationship between review valence and purchase intention (Beneke, de Sousa, Mbuyu, & Wickham, 2016; Kostyra et al., 2016; Lin & Xu, 2017), others add critical notes or moderating factors to it. Hu, Koh, and Reddy (2014) for instance confirm that review valence matters but find that this must be captured differently: the rating of a product in a review does not significantly affect purchase intention but review sentiment does, as this affects purchases. Other works look at differing effects for different products, as Cui et al. (2012) find that review valence has more of an effect on search products than on experience products. Also regarding the type of products, Sen and Lerman (2007) state that consumers care more about negative reviews when evaluating a utilitarian product, while they are more likely to ignore them in the case of hedonic products. Jeong and Koo (2015) add that negative reviews are more effective when they were written objectively while for positive reviews this does not matter.

Though there is overwhelming support for this relationship between review valence and purchase intention, other works argue that other aspects of reviews have an influence as well. Duan, Gu, and Whinston (2008) for example look at box office sales in the movie industry and find that not valence, but the sheer number of reviews matters for purchases. This first finding may be industry specific, but the second conclusion shows an awareness affect that customer reviews may also have. In fact, it confirms an earlier finding by Park et al. (2007) who show that the number of reviews a product has affects customer purchase intention in a virtual shopping mall setting. Kostyra et al. (2016) also stress the influence of review volume but find that it has a moderating effect through valence: a product with positive reviews is bought more when there is a high volume of these positive reviews. Finally, in addition to the number of reviews Park et al. (2007) find that the quality of a review matters for purchase intention, as they show a positive relationship between these two.

Online reviews thus help customers in the decision-making process and as such affect purchase behavior. This is relevant for managers to know as it implies that bad reviews can significantly harm their sales. However, the power of customer reviews goes beyond just sales. As mentioned in the introduction, online product reviews can actually decrease the role brand image has in this purchasing process (Kostyra et al., 2016). Since an organization's brand image is its most valuable asset, this impact of customer reviews is also relevant for managers (Aaker, 2009). The most popular work that looks at this effect is by Bambauer-Sachse and Mangold (2011). They use the term brand equity dilution as a possible effect of customer reviews. In general they find that negative customer reviews can hurt a firm's consumer-based brand equity as a whole. This thus again addresses the effect of a review's valence, which is something other works have done as well. Lin and Xu (2017) show for example that a review's valence has a significant positive effect on customer's brand attitudes after being exposed to them. Beneke et al. (2016) also find that negative reviews harm brand equity as a whole, but they add to this that this effect is larger when it involves a high-involvement product.

3.3 Drivers of online product review valence

Given these substantial effects that reviews have on brands and their sales, it is relevant for managers to know what drives their valence. This is however a topic that has been studied far less (Goes, Lin, & Au Yeung, 2014). In this next section we outline the works that have looked at potential drivers of review valence, based on which we then develop the hypotheses of this paper. We first look at drivers from the consumer perspective, and then outline what an e-commerce platform can do to influence review valence.

3.4 Review valence from the consumer's perspective

In theory, an online product review should be a reflection of a consumer's post purchase evaluation of that product (Li & Hitt, 2008; Moe & Schweidel, 2012; Moe & Trusov, 2011). This evaluation depends on the product that is sold, as for instance products that are more useful get more positive reviews (Moldovan, Goldenberg, & Chattopadhyay, 2011). Most of these product characteristics are not relevant for this study however, as we look at differences in review valence between the same products on different marketplaces. A factor that is relevant to consider is product price, as this is at least partly controlled by the seller of a product. Price matters, as a higher price for the same product means that consumers have lower utility from buying it, leading to a lower post purchase evaluation (Li & Hitt, 2010). Other than product-level characteristics, consumer characteristics matter as well, as a post purchase evaluation is subjective (Goes et al., 2014). An influential piece by Li and Hitt (2008) looks at this, as it analyzes self-selection in customer reviews. This bias, referred to in later work as the acquisition bias (Hu, Pavlou, & Zhang, 2017) exists because the first people to review a product are the early adopters who are usually fans of the product and brand, and thus have a more positive evaluation of the product. As such, products generally have inflated positive reviews in the beginning that decrease in positivity over time, but in general still remain higher than the actual product quality. This dynamic leads to the average review rating not reflecting the actual product quality.

This self-selection already shows that customer reviews are not just reflections of that product's quality and can therefore be different for the same product across marketplaces. Building on Li and Hitt (2008), there is a second stream of literature that investigates how reviews posted by others affect the valence of individuals writing reviews (Goes et al., 2014). Several works for instance find that as more reviews are posted, they tend to become more negative because consumers no longer find it necessary to contribute more positive reviews if these are already widely present (Godes & Silva José, 2012; Moe & Trusov, 2011; Wu & Huberman, 2008). However, the way individual reviewers react to previous reviews also differs. More frequent reviewers who consider themselves experts tend to post ratings that contrast that of the majority, while inexperienced posters tend to follow the already present sentiment (Moe & Schweidel, 2012) . Also, reviewers who are more popular tend to post more objective reviews that tend to be more negative and varied (Goes et al., 2014).

3.5 Review valence from the firm's perspective

In addition to factors on the consumer side, the company hosting the customer reviews can also exert influence on the valence of those reviews. This has also not been widely researched, but there are several works we can

draw on. Godes et al. (2005) provide a framework of four generic strategies that a firm has in managing what they refer to as social interactions, among which we can place online product reviews. These four strategies are for the firm to be an observer, moderator, mediator and/or participant in these social interactions. We now present the influences a firm can have on review valence that have been researched using this framework.

As an observer, a firm can use online product reviews but also other forms of eWOM to find out what consumers think of its brand or products (Godes et al., 2005). Given the valuable insights this can provide, a wide range of works exists that use text analytics applications to extract these insights from the enormous amount of eWOM data available (Rambocas & Pacheco, 2018). For instance, firms can use sentiment analysis to find out what consumers like and dislike about their brands and products by analyzing online products review or social media conversations (Ireland & Liu, 2018; Liu et al., 2017; Olagunju, Oyebode, & Orji, 2020). However, merely observing is not a strategy that firms can take to influence the valence in their product reviews.

The second strategy to managing reviews and their valence is for a firm to be a moderator of online product reviews, in which the first choice is to have a review section or not (Godes et al., 2005). As a moderator, the firm can take several steps that have an influence on the valence of reviews. The design of the venue in which these reviews are written and expressed are known to influence review valence, although it is unclear what exact design elements cause what effect (Smith, Fischer, & Yongjian, 2012). A specific step that a firm can consider to take is to let readers rate the helpfulness of the review or giving respected reviewers badges, increasing review credibility. With regard to valence these more helpful or expert reviews tend to be more negative, partly because reviewers who consider themselves experts tend to post reviews contrasting the (more positive) majority and partly because consumers rate more negative reviews as more helpful (Moe & Schweidel, 2012; Rosario, de Valck, & Sotgiu, 2020; Schuckert, Liu, & Law, 2015). Another specific step a firm can take to influence reviews is to incentivize consumers to write them after a purchase, for instance by offering them a discount coupon or even a monetary reward for writing one (Garnefeld, Helm, & Grötschel, 2020). This is known to increase the number of reviews written, but it does not necessarily positively influence the valence of those reviews. The valence is positive when product satisfaction is high, but incentives may also lead to more negative reviews if product satisfaction is low (Garnefeld et al., 2020). However, firms can influence the valence of these incentivized reviews by including examples and recommendations for review to be written, as this is known to influence the review writer (Kim, Naylor, Sivadas, & Sugumaran, 2016). Finally, brand loyalty can be a deciding factor in determining review valence. If firms want consumers to write positive reviews about their brands or products online, they must actively encourage them to do so by activating their ties to the brand (Eelen, Özturan, & Verlegh, 2017). This is where a good CRM system and 'owning' the customer relationship matter, as it can give a firm the power to influence the valence of the reviews on its website.

The third strategy for a firm is to be the mediator of online social interaction. This lies very closely to the role of moderator, but in this strategy the firm actively takes control of to whom the information is provided (Godes et al., 2005). However, this applies more to providing references to potential clients and is not applicable to online product reviews (Godes et al., 2005).

Staying with this more active role of the firm, the final strategy is active participation of the firm

(Godes et al., 2005). Outside of the online product review context this could consist of a company getting involved in forum or social media conversation about her products (Homburg et al., 2015). Within the context of online products reviews a well-researched strategy is review manipulation, which may involve posting positive reviews under a different name or deleting unfavorable reviews (Rosario et al., 2020). This is obviously unethical, but around 1 in 10 online product reviews is known to be fake (Hu, Bose, Koh, & Liu, 2012). Interestingly, there are scenario's where each firm engaging in these practices does not harm consumers, but in the end it can lead to companies getting into 'rat races' where they all spend resources to manipulate reviews without any gain, as everyone else does it as well (Dellarocas, 2006). Manipulation of reviews especially backfires when consumers are able to detect it, which is most likely to happen in case of fake reviews being posted. Deleting unfavorable reviews is a more viable strategy, but it is still unethical and must be done with restraint (Zhuang, Cui, & Peng, 2018). In this paper we cannot credibly say that either the manufacturer or the seller engages in review manipulation. However, something that is interesting to note is the fact that a manufacturer selling its products through its own direct channel has more of an incentive to artificially improve its review sentiment than a third party marketplace like Amazon: the third party marketplace also has other products in that category from other brands that consumers can consider. In fact, Amazon has guidelines in place making sure that vendors on the platform (among which manufacturers) do not discourage consumers from posting negative reviews about their products in the Amazon review section (Rosario et al., 2020).

3.6 Conceptual framework and hypotheses

Building on the drivers of customer reviews above we now present the conceptual framework and hypotheses. This framework is visualized below in Figure 1. We first introduce the main effect studied, followed by the moderating effects. Finally we also present other direct effects on review valence, which we need to control for in our analysis.

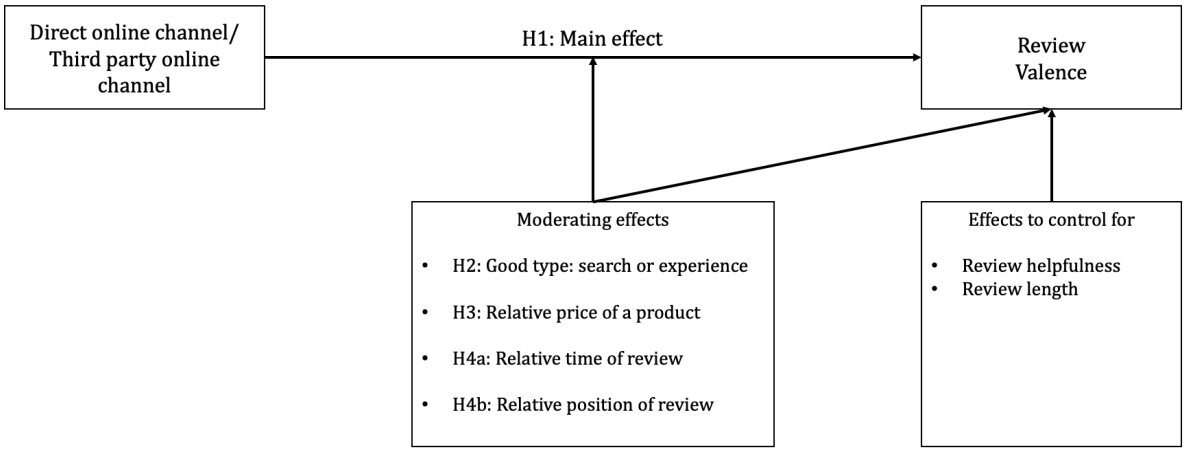

Figure 1: The conceptual framework of this paper

3.6.1 Main effect

Type of marketplace

The main effect studied here is the relationship between the type of marketplace and review valence. We hypothesize that the valence of reviews of the same product is more positive in the direct channel. We base this hypothesis on two mechanisms from the theory discussed above. First, we think that the self-selection bias of reviewers presented by Li and Hitt (2008) is stronger for direct channels. This theory expects that reviews will be more positive than a product's quality suggests, as avid consumers of that brand will buy the products first and post more positive reviews from the start (Hu et al., 2017). We expect this effect to be stronger in the direct channel, as we expect consumers who purchase their product in a brand's owned store to be more loyal to that brand, leading to higher initial reviews that then stay higher over time. Closely connected to this notion of more brand loyalty, we expect higher valence in the direct channel's product reviews because of more effective elicitation of positive reviews after purchase. Neither marketplace can offer incentives to leave positive reviews but it is possible to encourage consumers to write a review (Garnefeld et al., 2020). However, when consumers are more brand loyal it is easier for a marketer to encourage her customers to leave positive reviews (Eelen et al., 2017). Therefore we think that the direct channel can better elicit positive reviews. In summary our first hypothesis is:

Hypothesis 1 (H1): *The valence of reviews on the same product is more positive in direct online sales channels compared to third party online sales channels*

3.6.2 Moderating effects

For the main effect studied we hypothesize that reviews in the direct online channel have more positive reviews compared to those in a third party online channel. To provide more depth in our analysis of this relationship, we also want to look for moderating effects in this relationship. These moderators and corresponding hypotheses are presented below.

Good type: search or experience

The first moderating effect we expect to find in the relationship between review valence and marketplace type is whether a review is written about a search or an experience good. More formally, search goods are goods whose quality can be evaluated objectively using factual information about the products' attributes that is available before purchase (Mudambi & Schuff, 2010). On the other hand, experience goods are goods whose quality can be evaluated based on subjective attributes that are hard to access and require feeling or experiencing the product. This moderator was not mentioned in our study on drivers of review valence but in our review of the consequences of online product reviews for firms. Well-cited work by Cui et al. (2012) finds that the effect of valence on a good's sales is stronger for search goods compared to experience goods. They argue that product reviews are among those easy to access attributes used to evaluate search goods' quality, and as a result play a larger role in the purchase of search goods compared to experience goods. We include this distinction between goods as a moderator not only to make a new contribution to the literature but also to give managers more useful insights. As the effect of review valence on final sales is known to be different, it is also relevant for managers to know how this valence differs over marketplaces depending on the good type.

We hypothesize that the difference between review valence between a first party online sales channel and a third party online sales channel is larger for experience goods. This is because we expect the self-selection

bias central to the mechanism supporting our first hypothesis to be even stronger for experience goods. We draw this hypothesis by looking more in depth at the self-selection bias and the underlying framework presented by Li and Hitt (2008). As discussed above, they find that review valence tends to be higher than the actual product quality would suggest due to early adopters who are fans of the product or brand buying and reviewing the product early on, after which the valence decreases but still stays higher than the products' quality. In their work they actually study this mechanism in the context of an experience good, which has both search and experience attributes. They argue that in the case of the positive self-selection bias they find evidence for, these two are positively correlated: early adopters purchase a product early as they likely appreciate the search attributes they observe beforehand, and then also have a more positive evaluation of the more subjective experience attributes after purchase. We expect that as a product has relatively more experience attributes, this mechanism behind the self-selection becomes stronger. This means that the early reviews will be more positive for experience goods, and make the difference between the valence in the two different marketplaces even larger. For search goods on the other hand, there are fewer or no experience components. A consumers' purchase choice and subsequent evaluation of the product thus revolves more around their objective evaluation of the product. Reviewers on the first party online sales channel may now still be more brand loyal than those in the third party online sales channel, but their evaluation of the product will be based less on subjective evaluations of the experience attributes and thus the difference in valence between their reviews and those on third party channels will be lower. In sum our hypothesis for this effect is:

Hypothesis 2 (H2): *The difference in review valence for the same product between a direct online sales channel and a third party online sales channel is larger for reviews about experience goods compared to reviews about search goods.*

Relative price

The second moderating effect we study looks at the relative price of a product. This addresses the difference in price between both marketplaces, not the absolute price of a product. The absolute price should in our application not affect the difference in valence, as an expensive product will be expensive in both marketplaces. We account for this difference in prices because review valence partly depends on consumer's post-purchase evaluation of a product (Moe & Schweidel, 2012; Moe & Trusov, 2011). Work by Li and Hitt (2010) argues that this evaluation reflects a product's value to a consumer rather than its quality, and that when a product's price is increased this means that review valence becomes more negative as the product's value decreases. In our application we look at the same product sold in different marketplaces that may charge different prices. For our main effect we think that reviews in the direct channel are on average more positive because the customers purchasing in that channel value the brand or product more. Adding the price difference to this equation, we expect these more loyal customers to be less price sensitive than the more critical customers in the third party channel. This means that when the price of a product is relatively higher (lower) in the third party channel, the value or utility derived from that product by those customers is lower (higher) than if the same product were bought in the direct channel where customers care less about this price difference. This leads to the hypothesis:

Hypothesis 3 (H3): *The difference in review valence for the same product between a direct online sales channel and a third party online sales channel is smaller (larger) when the price of that product is higher (lower) in the direct online sales channel, compared to the third party online sales channel.*

Relative timing of a review

The third moderating effect addresses the position of a review relative to the other reviews on a specific product. The choice of this moderator originates from two findings by earlier works: a temporal effect found by Li and Hitt (2008) and a sequential effect found by Wu and Huberman (2008). Starting with Li and Hitt (2008), they find that the self-selection bias we expect to drive positive review valence in the first party online sales mostly takes place in early reviews. As such, we expect that reviews on a product that were written at a later point in time are less positive, as they approach the actual product quality and thus also the valence levels observed in the third party sales channel. Work by Wu and Huberman (2008) also states that there should be a declining trend in review valence, but this valence is expected to become less positive as more reviews written rather than when time passes. They argue that consumers are motivated to write reviews because they want to help others in their decision-making process. If there are already many positive reviews present, as our first hypothesis expects to be the case especially in the direct online channel, consumers who like the product no longer feel the need to express this in a positive review because their review would not add any new information for subsequent consumers. As such only those who have contrasting, negative opinions about a product will write a review, leading to a negative trend as more reviews are written.

Both these theories thus expect a negative trend after the positive early reviews that we expect to see especially in the direct channel. As a result we expect that the relative timing of a review moderates the difference in valence between the two types of marketplaces, as we expect it to be smaller when a review was posted relatively later. However, each work addresses a different meaning of relative timing, as Li and Hitt (2008) address the time a review was written and Wu and Huberman (2008) address the position in terms of how many reviews precede a review. Work by Godes and Silva José (2012) integrates both these effects into a model and finds evidence for both, but most importantly stresses that future work should incorporate both these effects in their analyses. We follow this recommendation and include two moderating effects in our framework. This means that we separate this moderating effect into two hypotheses, 4a and 4b.

Hypothesis 4a (H4a): *The difference in review valence for the same product between a direct online sales channel and a third party online sales channel is larger for reviews written earlier in time compared to reviews written later in time*

Hypothesis 4b (H4b): *The difference in review valence for the same product between a direct online sales channel and a third party online sales channel will be larger for reviews that have few reviews preceding them compared to reviews that have many reviews preceding them*

3.6.3 Effects to control for

Review helpfulness and review length

We expect that review helpfulness as perceived by users is important to include as a control variable because it is a useful for proxy for actual drivers of review valence. We base this on the drivers of review valence identified above. First, there exists empiric proof that consumers tend to rate more negative reviews as more helpful (Moe & Schweidel, 2012; Rosario et al., 2020; Schuckert et al., 2015). However, this is not an effect to control for as these helpfulness votes take place after the reviews was written and thus the valence was determined. A better reason to include this as an effect to control for originates from the social dynamics of review writing, as we know that more experienced and/or more popular reviewers tend to write more negative reviews (Goes et al., 2014; Moe & Schweidel, 2012). The difficulty here though is that due to a lack of data on individual reviewers we cannot identify expert or popular reviewers. Including review helpfulness

solves for this, as we use it as a proxy for a reviewer's expertise and/or popularity: reviews that have high helpfulness as perceived by other users are likely written by experts and/or popular reviews, who tend to write more negative reviews.

Another direct effect on review valence to control for is a review's length. Homburg et al. (2015) show that longer reviews on average have lower valence, but do not give a theoretical reason for it, as it is only a control variable in their study. One reason why this effect takes place may lie in the motivation theory argument by Wu and Huberman (2008): consumers will only take the effort to write a review when they have a contrasting opinion. As longer reviews are more costly to write, we might expect their valence to be more contrasting. As a result we would observe longer reviews that are more negative, contrasting the overall positive valence observed in most online product review sections (Chevalier & Mayzlin, 2006; Ireland & Liu, 2018).

Table 1 below summarizes the hypotheses and effects to control for.

Main effect	Underlying mechanis(m)	Hypothesis
Type of Marketplace	Self-selection bias (Hu et al., 2014; Li & Hitt, 2008) Elicitation of positive reviews (Eelen et al., 2017; Garnefeld et al., 2020)	H1: The valence of reviews on the same product is more positive in direct online sales channels compared to third party online sales channels
Moderating effects	Underlying mechanis(m)	Hypothesis
Good Type: Search or Experience	Self-selection bias is stronger among experience goods (Li & Hitt, 2008; Mudambi & Schuff, 2010)	H2: The difference in review valence for the same product between a direct online sales channel and a third party online sales channel is larger for reviews about experience goods compared to reviews about search goods.
Relative Price	Post-purchase evaluation (Li & Hitt, 2010)	H3: The difference in review valence for the same product between a direct online sales channel and a third party online sales channel is smaller (larger) when the price of that product is higher (lower) in the direct online sales channel, compared to the third party online sales channel.
Relative timing of a review	Self-selection bias is mostly present in early reviews (Godes & Silva José, 2012; Li & Hitt, 2008)	H4a: The difference in review valence for the same product between a direct online sales channel and a third party online sales channel is larger for reviews written earlier in time compared to reviews written later in time
Relative position of a review	Consumers only post reviews when they contrast the already present positive reviews (Godes & Silva José, 2012; Wu & Haberman, 2008)	H4b: The difference in review valence for the same product between a direct online sales channel and a third party online sales channel is larger for reviews that have few reviews preceding them compared to reviews that have many reviews preceding them
Effects to control for	Underlying mechanis(m)	Hypothesis
Review Helpfulness	Negative reviews are rated as more helpful (Moe & Schweidel, 2012; Rosario et al., 2020) Expert reviewers write more negative reviews (Moe & Trusov, 2011) Popular reviewers write more negative reviews (Goes et al., 2014)	N/A
Review Length	Empiric findings by Homburg et al. (2015) Motivation theory (Wu & Huberman, 2008)	N/A

Table 1: The effects outlined in this section and their corresponding hypotheses

4 Data

In this section we present the data used in this paper. We first discuss the source, then present our variables and finally show some simple analyses of it.

4.1 Data source

To test the hypotheses presented in the framework above, our study requires data that includes online product reviews for the same products of a brand, in two different marketplaces: one direct online channel owned by the brand and one third party online sales channel. Also, we need two types of goods: one search and one experience good. Starting with the good types, our search good is represented by Digital Single-Lens Reflex (DSLR) cameras. Digital cameras, part of the broader category of consumer electronics, have been used to represent a search good in many previous studies (Cui et al., 2012; Luan, Yao, Zhao, & Liu, 2016). As we study one brand's products over two marketplaces, we also require a choice of one brand among these cameras. The brand we choose for the search good is Nikon. The first reason for it is that it is one of the largest players in the market only behind Canon, making it a company with substantial brand value (Statista, 2020). However, there is also a practical reason as Nikon has a direct online channel with reviews that does not block web scraping. For our experience good we use Asics running shoes. Running shoes, but also sportswear in general, have also been used before to represent experience goods (Luan et al., 2016). Also, running shoes are intuitively experience goods: one does not know the real quality until he/she goes for a run on them. Asics, like Nikon, is a large player in the running industry, and thus also has a brand that can be harmed by online reviews (Richter, 2021).

For the marketplaces, the first party marketplaces are Nikon's and Asics's respective stores on their own websites. For the third party marketplace we look at Amazon, which we do for two reasons. First, Amazon is the undisputed market leader in e-commerce and thus the most likely third party marketplace consumers turn to for the average good (Markinblog, 2021). Second, Amazon is heralded for its online product reviews which have been used for many studies before (Chevalier & Mayzlin, 2006; Ireland & Liu, 2018; Olagunju et al., 2020). Note that, for consistency purposes, we use the US website for each of the marketplaces. This is useful because we need reviews written in English and because the US is the largest English speaking market, making for the largest variety in products and reviews within each product group we look at.

To obtain this data, we make use of web scraping. This technique allows us to systematically obtain the online product reviews using only the product page URLs as input. Scraping Amazon was especially straightforward, as we could use the `rvest` package in R to scrape this data (Wickham, 2020). The only caveat here is that we can only scrape reviews from US customers due to technical reasons, but the vast majority of reviews is written by US customers anyways. For both Nikon's and Asics's brand stores, this process was less straightforward because these websites use JavaScript. As a result, the `rvest` package could not work with them. To scrape these reviews, we use the free-to-use scraping tool Parsehub. Besides requiring no coding skills, this tool has the advantage that it can read JavaScript websites, making it ideal for our task.

4.2 Preparing the data

For each of the four combinations of good type and marketplace we scrape two datasets: one general dataset containing an overview of the specific products in that category and marketplace, and one containing the actual reviews for all these products. We do this because we first need to obtain the specific product page URLs so that we can then revisit these to obtain the reviews. After scraping and before preparing the data we have 1122 reviews on 17 products from Nikon's online store, 9348 reviews on 26 products from Nikon on Amazon, 7224 reviews on 167 products on Asics products from Asics's own store and 30,339 reviews on 44 products from Asics on Amazon. First, we see that Amazon has more reviews than the direct channels, which makes sense given that Amazon is renowned for its review system (Ireland & Liu, 2018). Also, we see that Asics has many more shoes than Amazon. This is because Asics has many special versions of the same shoe, while Amazon mostly has the most popular models. The first step of preparing the data involves combining these reviews with the general product information, such that each review on a specific product has the relevant product-level variables accompanying it. Then, we format each variable such that it becomes useful for further analysis, as we for instance convert a scraped value of 'Rated 4 out of 5' to a numeric star rating of 4.

After cleaning our scraped review data, the next step is combining the separate datasets. Because our aim is to compare review valence for the same product's reviews across marketplaces, our final dataset only includes reviews on products that are sold and have reviews in both. To find this overlap, we use the product names, which after some formatting can be used for finding matching products between the direct channel and Amazon. Based on these product names we gave each product a unique ID for that name, meaning that products with the same name across a direct channel and Amazon get the same ID number. When then merging the datasets, we only keep the reviews that have IDs occurring in both marketplaces' datasets. These IDs will later on be used to account for product-specific effects, to be explained in the methodology section.

Unfortunately we had no access to a unique product code to identify the same product across marketplaces, so this strategy is not fully waterproof. For the Asics store, which has separate product and review pages for different color styles of the same shoe, this results in a lower number of reviews on one product page, spread out over multiple different product pages. Meanwhile Amazon places all these different styles of that same shoe on one page along with one review page. Using our merging strategy, we treat all these different pages of the same shoe as one, and match the reviews to reviews on Amazon using the unique ID. This method allows us to control for most of the product-level differences such as general quality that may influence review valence.

The final dataset consists of 18,979 observations in total, of which 944 reviews on 10 products from Nikon's store, 2334 reviews on 10 Nikon products from Amazon, 7489 reviews on 26 products from Asics' store and 8212 reviews on 26 Asics products on Amazon.com. Note that especially for Asics the number of products has dropped by a lot: this is the case because Asics running shoes have several special editions of the same shoes, which usually have the same review pages or are generalized when combining the two datasets we scraped for Asics.

4.3 Variables

Having prepared the data, we can operationalize the effects identified in the framework of the previous section. We present these variables below, starting with the dependent variable, followed by our independent variables capturing the main effect, moderating effects and effects to control for.

4.3.1 Dependent variable

The dependent variable used in this paper is the valence of an online review. As explained before, we research how this differs across sales channels for the same product to provide empirical evidence that this is a factor a manufacturer should consider in choosing an online sales channel. In our empirical setting, we capture a review's valence using two different operationalizations. First we use a review's accompanying star rating as a representation of its valence, as many studies have done this before and have found that variation in this variable indeed have implications for a company's sales and brand (Chevalier & Mayzlin, 2006; Cui et al., 2012; Zhu & Zhang, 2010). Second, we do the same analysis with review sentiment as the dependent variable. Some studies argue that review sentiment and star rating similarly capture review valence (Ireland & Liu, 2018), but Hu et al. (2014) find that a review's valence affects sales through review sentiment and not through review star rating. Moreover, other studies simply use review sentiment as their dependent variable capturing review valence (Homburg et al., 2015; Liu et al., 2017; Olagunju et al., 2020).

To capture this sentiment, we use the `polarity` algorithm, which is part of the `qdap` package in R (Rinker, 2020). This algorithm uses a dictionary approach to classify a review's sentiment, and has been used for a wide range of application in previous works (Bhavaraju, Beyney, & Nicholson, 2019; Fernandes, Moro, Cortez, Batista, & Ribeiro, 2021; Xu, 2018). In each review, it looks for words that indicate a sentiment, which it recognizes because it works with an underlying dictionary that contains over 6,000 polarity words that either indicates a positive or negative sentiment. For each positive word it finds in a review that review gets +1 point and for each negative word is gets -1 point. It also accounts for negations which it also recognizes through another dictionary containing negation terms. For each negation term that occurs either four words before a word indicating sentiment or two words after it, the sign of the 1 point is flipped (e.g. 'not good' gets -1). Furthermore, it accounts for amplification words such as 'very' or 'really'. Using the same window as for negation words, the algorithm scales the points awarded by (1+0.8), meaning that 'really good' gets 1.8 and 'really bad' gets -1.8. One special case is the combination of an amplification and a negation word, for instance 'not very good'. Now the scaling is not by (1+0.8) but by (1 - 0.8), meaning that this example gets a score of 0.2. Finally, it sums all these points for one review and then divides it by the square root of the number of words in a review. This is to account for the density of polarity words, as a short review with three positive words in one sentence is very positive and should not get the same sentiment score as a long review consisting of three sentences with one polarity word in each. This is thus different to the idea behind controlling for review length we discussed in the section above. The final result of this algorithm is a score for reviews below zero with negative sentiment and a score above zero for positive sentiment.

4.3.2 Independent variables

Main effect: Type of Marketplace

As discussed in the conceptual framework, our main effect studied is how valence differs over the two marketplaces. We operationalize this in our model through a dummy variable *Type of Marketplace* that has value 0 for reviews on the direct online sales channel and value 1 for reviews on Amazon. We manually coded this variable when merging the datasets as described above.

Moderating effect: Good type

The first moderator we add to our model is the type of good. Like the marketplace, we operationalize this through the dummy variable *Good Type*, where the search good, Nikon cameras, has value 0 and the experience good, Asics running shoes, has value 1. This variable was also manually coded when merging the datasets.

Moderating effect: Relative Price

The second moderating effect in our framework addresses the *Relative Price* of a product. The *Relative* in *Relative Price* means that we look at the difference between a product's price in US dollars between the two marketplaces. We use the direct channel as a reference here, meaning that when the price of its product is lower (higher) on Amazon than on the direct channel, a review has a negative (positive) value for this variable. Obtaining the prices of products proved to be a challenge, as Asics gave promotions on prices and Amazon has variable pricing. For Asics, we took the non-promotion price if there were multiple versions of the same model that were also sold for that non-promotion price, as this indicates a temporary nature of that price. For Amazon, we selected size 9 of the default color on each product's page. This is for consistency purposes, but may make our results regarding this variable unstable.

Moderating effect: Dayssince and Number of Preceding Reviews

The final moderating effect addresses the timing of a review relative to earlier reviews on that product in a specific marketplace. Following a recommendation by Godes and Silva José (2012), we separate this effect in two variables. First, we look at a temporal effect by creating the variable *Dayssince*, which indicates the number of days between a certain review and the first review on that product in a marketplace. Second, we look at a sequential effect through the variable *Number of Preceding Reviews*, which captures how many reviews have preceded that review on a specific good in a marketplace. Both these variables were coded manually in the early stage of the data preparation process, as we still had all reviews present at that time.

Control: Review Helpfulness and Review Length

The control variables we include, next to the moderators above which we also add separately to the model, are *Review Helpfulness* and *Review Length*. The helpfulness of a review is a variable that we could scrape together with the individual reviews, as all marketplaces have an option for consumers to vote on whether or not they find a review helpful. For Amazon this is fairly straightforward, as each review has a number of votes starting at 0. For Nikon's and Asics's online stores this variable looks a bit different, as consumers can vote 'yes' or 'no' for helpfulness, meaning that each review has two vote counts. To make these numbers comparable with Amazon's, we first recoded this variable for Nikon and Asics by subtracting the 'no' votes from the 'yes' votes. This means that among the non-Amazon reviews we also got reviews with negative scores. This is an issue, as consumers on Amazon never had the chance to downvote a review, meaning that

a 0 on Amazon means that no one felt the need to vote or that no one found the review helpful. To make this scaling consistent over all marketplaces, we set all negative helpfulness scores for non-Amazon reviews to 0 as well, as it now also indicates that consumers either did not find the review helpful or had no reason to vote. Also, reviews from different marketplaces or on different products have very different scales in terms of what is a good review or not, for example because simply more consumers visit one store and thus tend to vote more in general. This can result in a review with 5 'yes' votes in the Nikon store to be actually more helpful than a review on a very popular Asics running shoe with 20 helpful votes in the Amazon store, simply because less consumers buy a Nikon camera or visit the Nikon store. To solve for this, we standardize the variable *Review Helpfulness* relative to each marketplace and good. We do this by dividing each value by the standard deviation of *Review Helpfulness* for the reviews in that specific marketplace-good type combination. This is different to normal standardizing where one should first subtract the mean and then divide by the standard deviation, but we do not do this to keep helpfulness scores of 0, which the majority of reviews have, indeed equal to 0 and equal across marketplaces and good types.

The final control variable *Review Length* is more straightforward to make. It is a numeric variable that captures the number of characters in a review. It was manually created in the first steps of the data collection progress, as we want the number of characters in the raw, pre-cleaned reviews, since this is the length of the review as the original writer wrote it.

4.4 Descriptive statistics

Now that the final dataset is ready, we can explore it before moving to the empirical analysis of this paper. Table 2 below show the descriptive statistics for all variables we summarized above. This overview is already divided among the marketplaces, allowing us to already get an idea of how our variables may differ between them. We also do this for the type of good (Nikon cameras vs Asics running shoes) later in this section.

	Direct Channel					Third Party Channel (Amazon)				
Variable	Obs	Mean	Std. Dev.	Max	Min	Obs	Mean	Std. Dev.	Max	Min
Dependent										
Star Rating	8433	4.425	1.099	5	1	10450	4.440	1.157	5	1
Sentiment Score	8433	0.557	0.545	3.6	-1.878	10450	0.515	0.541	7.788	-1.273
Independent										
Good Type	8433	0.888	0.315	1	0	10450	0.779	0.415	1	0
Relative Price	8433	18.735	95.919	697	-371.95	10450	25.196	139.8	697	-371.95
Dayssince	8433	294.342	290.038	2384	0	10450	402.568	303.667	6030	0
Number of Preceding Reviews	8433	86.12	92.59	458	0	10450	1053.535	1316.973	6030	0
Review Helpfulness	8433	0.345	1	15.311	0	10450	0.132	1	43.466	0
Review Length	8433	203.234	288.999	9289	11	10450	239.752	557.864	18883	1

Table 2: Descriptive statistics of the variables used in this paper, separated by marketplace type

The first thing to note here is the number of observations, as there are more reviews from Amazon than from the direct online channels. When then looking at the dependendent variables, we see that the first operationalization *Star Rating* has a mean around 4.4 for both marketplaces. Considering that this star rating is a five point scale, this implies that the average review is very positive. This was expected for Amazon, but it seems that it also the case for Nikon's and Asics's brand stores (Chevalier & Mayzlin, 2006). For *Sentiment Score*, the other operationalization of review valence, we see a mean score of 0.557 in the

direct channels and 0.515 on Amazon. Both are on average positive, but there seems to be at least a small difference between the marketplaces now. Also interesting to note is the fact that the most positive review on Amazon is a lot more positive, at a score of 7,788 versus the maximum sentiment score of 3.6 on the direct channels. Moving to the independent variables, we see similar statistics for *Good Type*, as between 78-89% of the reviews in both marketplaces are written on our experience good, the Asics running shoes. For the final moderator, *Relative Price*, we see that Amazon tends to be more expensive, as the mean of this variable is positive for both. Note that these means are different across marketplaces, even though the value should be the same for each product regardless of marketplace. This is because we have more reviews on a certain product is not the same across marketplaces, meaning that Amazon probably has many reviews on a product that has a higher price on Amazon. Also, there are large differences in prices as shown by a maximum value of 697 and minimum value of -371.95. This takes place among the Nikon cameras, which can cost above 2000 dollars. The statistics for review timing (*Dayssince*) are very similar, but for *Number of Preceding Reviews* this is not the case. This variable has a mean of 86.12 on the direct channels, implying that the average product has 86 reviews here. On Amazon this is way higher, as the mean of 1053.535 implies that there are on average 1054 reviews on a product on Amazon*. For *Review Helpfulness* we do not see very large differences, which makes sense given that this variable has already been standardized. Finally review length also shows similar numbers across both marketplaces, the only thing that is very different here is the maximum length of a review, which is twice as high on Amazon.

	Search Good (Nikon Cameras)					Experience Good (Asics Running Shoes)				
Variable	Obs	Mean	Std. Dev.	Max	Min	Obs	Mean	Std. Dev.	Max	Min
Dependent										
Star Rating	3278	4.569	1.049	5	1	10914	4.415	1.159	5	1
Sentiment Score	3278	0.521	0.485	7.788	-1.124	10914	0.531	0.556	4.246	-1.878
Independent										
Type of Marketplace	3278	0.712	0.453	1	0	10914	0.744	0.437	1	0
Relative Price	3278	52.135	248.194	697	-371.95	10914	11.991	12.149	80	-18.11
Dayssince	3278	589.619	499.539	2709	0	10914	342.031	197.796	1046	0
Number of Preceding Reviews	3278	1950.036	1972.429	6030	0	10914	465.395	477.433	2059	0
Review Helpfulness	3278	0.215	1.001	28.162	0	10914	0.169	1.004	43.282	0
Review Length	3278	542.28	983.682	18883	1	10914	156.496	182.083	3376	1

Table 3: Descriptive statistics of the variables used in this paper, separated by good type

Looking at the same statistics in Table 3, now separated by the type of good, we see that each good has similar figures for both dependent variables. Interestingly though, the cameras have a slightly higher average star rating whereas the running shoes have a slightly higher average sentiment score. For the independent variables, we now have *Type of Marketplace* where we had *Good Type* and the mean of this variable shows that 71% of the camera reviews and 74% of the running shoe reviews are from Amazon. For the relative prices, we see that the Nikon cameras have higher values for all statistics, which makes sense because they are more expensive than the running shoes. For both *Dayssince* and *Number of Preceding Reviews* Nikon cameras have a higher mean, indicating that these products receive reviews for a longer period of time and receive more reviews in general. The mean and maximum of *Review Length* add to this that these reviews also tend to be longer.

*This difference in number of reviews per product occurs not only because people write more reviews on Amazon, but also because of the way product pages are organized. For Asics shoes, different styles of the same shoe have one review page each, while Amazon organizes all reviews for these different shoe styles on the same page.

As a final exploration of our data we show the correlation between each of the variables below. It is important to check for this because we do not want our results to be distorted by the presence of multicollinearity between variables. The matrix shown in Figure 2 below shows this.

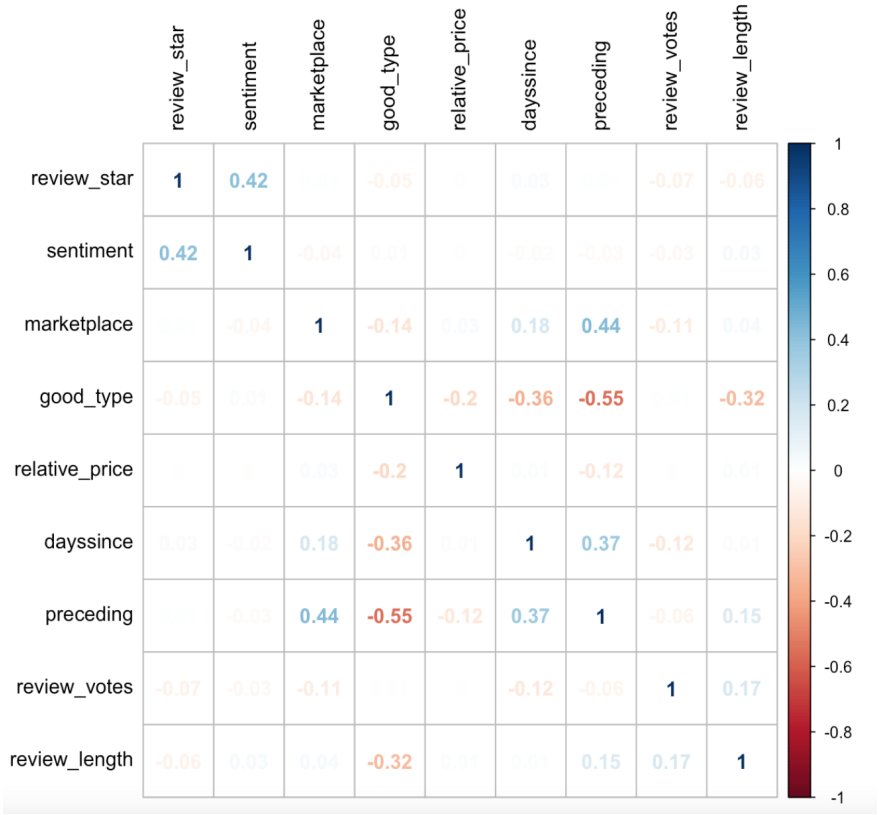

Figure 2: Correlation matrix including the variables used in this paper

Starting in the top left corner, we see that both operationalizations of valence review_star (*Star Rating*) and *Sentiment* have a correlation coefficient of 0.42. One would expect them to be correlated, but interestingly this correlation is not close to 1, implying that they do not capture valence in exactly the same way. Moving on, we see no substantial correlation between marketplace and the dependent variable. The next noteworthy correlation is between preceding (*Number of Preceding Reviews*) and marketplace. This confirms what we already saw before: the number of reviews on one product differs a lot between the two channels. The strongest correlation present here is a negative correlation coefficient of -0.55 between preceding and good type. This implies that the number of reviews also differs a lot for both good types. Finally we see a very expectable correlation coefficient of 0.37 between preceding and dayssince, showing that products that have been receiving reviews for a longer time also receive more reviews.

5 Methodology

In this section we discuss the models used to test this paper's hypotheses. We have to develop these in such a way that we can deal with two important features of our data. The first of these features is the fact

that our data contains variables at different levels: we want to explain an individual review's valence using a higher level variable about the marketplace it is sold in, while controlling for product- and review-level factors. Second, we operationalize this valence in two ways requiring a different type of regression for each. We start this section by explaining how we deal with the different levels, as this is something that applies to the regressions for both marketplaces.

5.1 Multilevel modeling

When building an empiric model, it is important to recognize how the data used is structured. In our case, we have variables measured at three different levels, which we all use to explain an individual review's valence. These levels are, from low (Level 1) to high (Level 3): the individual review, marketplace and product levels. Table 4 below summarizes what variables belong to which level. If we were to put all of these variables in a regular Ordinary Least Squares (OLS) regression, this would become problematic (Petersen, 2009). The reason for this is the fact that OLS regression assumes that the error terms are independent. This assumption is violated with our data structure, as for instance all reviews on a certain product will have the same values for all product-level variables. As a result the OLS model will underestimate the standard errors of the higher level variables (i.e. marketplace or product level), which can make their effect on the lower level dependent variable (review valence) seem significant while it actually is not (Moulton, 1990).

Level 1: Review (i)	Level 2: Marketplace (j)	Level 3: Product (k)
Both dependent variables	*Type of Marketplace*	*Good Type*
Dayssince		*Relative Price*
Number of Preceding Reviews		
Review Helpfulness		
Review Length		

Table 4: Variables used and their corresponding level

There are several ways to solve for this issue. Work by Cheah (2009) replicates earlier work by Moulton (1990) and shows that the best way to do this is by using multilevel models, as this is the least likely to underestimate standard errors. It empirically tests how often each method to deal with nested data wrongly reject the null hypothesis that there is no significant relationship, and finds that multilevel regression models perform better than for instance models with clustered standard errors. This is why we opt to use this method as well. Among multilevel models, we choose to use a random intercept model. In the random intercept model, we model the relationship between the dependent variable at the lowest level (review) and the independent variables at the lowest level, while we let the higher level variables determine the intercept of that relation. This means that the valence of reviews with the same review-level features can vary for reviews on different products or on different marketplaces, but the relation between those review level independent variables and the dependent variable (i.e. a lower level variable's coefficient) is the same across reviews from different marketplaces or products.

Using the variables shown in the data section, we now illustrate our proposed model below. Note that the focus is here on the multilevel nature of our model, it will be adjusted to incorporate each operationalization

of the dependent variable later on in this section. Equation 1 below shows the model at its lowest level, the review level.

$$\begin{aligned} Review\,Valence_{ijk} &= \pi_{0jk} + \pi_1 * Dayssince_{ijk} + \pi_2 * Number\,of\,Preceding\,Reviews_{ijk} \\ &+ \pi_3 * Review\,Helpfulness_{ijk} + \pi_4 * Review\,Length_{ijk} + \pi_5 * Type\,of\,Marketplace_{jk} * Dayssince_{ijk} \\ &+ \pi_6 * Type\,of\,Marketplace_{jk} * Number\,of\,Preceding\,Reviews_{ijk} + v_{ijk} \end{aligned} \quad (1)$$

Here $Review\,Valence_{ijk}$ indicates the valence of review i posted in marketplace j where product k is sold, which can be *Star Rating* or *Sentiment Score*. The intercept π_{0jk} is affected by higher level variables at the marketplace level j and product level k, which we elaborate on below. Moreover we see the review level variables $Dayssince_{ijk}$, $Number\,of\,Preceding\,Reviews_{ijk}$, $Review\,Helpfulness_{ijk}$ and $Review\,Length_{ijk}$, which all apply to review i in marketplace j on product k. The final two parameters in this part of the model address the moderating effect of $Dayssince_{ijk}$ and $Number\,of\,Preceding\,Reviews_{ijk}$ on the main effect of $Type\,of\,Marketplace_{jk}$ respectively. As these interactions include variables from different levels, they are called cross-level interactions. These are usually used to evaluate how a higher level variable affects the effect of a lower level independent variable on the lower level dependent variable, but they are symmetric in nature (Aguinis, Gottfredson, & Culpepper, 2013). As such we use them the other way around to find out how a lower level moderator affects the relationship between the dependent variable and level 2 variable of interest, as this is what our conceptual framework expects. Finally we see the error term v_{ijk}, which is the error term for this level of variables.

Going back to the intercept, this is affected by marketplace- (j) and product (k) level variables. More formally, it is equal to Equation 2 below.

$$\begin{aligned} \pi_{0jk} &= \delta_{0k} + \delta_1 * Type\,of\,Marketplace_{jk} + \delta_2 * Type\,of\,Marketplace_{jk} * Good\,Type_k \\ &+ \delta_2 * Type\,of\,Marketplace_{jk} * Relative\,Price_k + e_{jk} \end{aligned} \quad (2)$$

Here, the intercept of the lowest level model, π_{0jk}, also has its own intercept that is affected by variables at the highest level product (k), represented by δ_{0k}. Also, we see that this intercept is affected by marketplace level variable $Type\,of\,Marketplace_{jk}$, a dummy indicating whether marketplace j where product k is sold is owned by the manufacturer or not. Here j has three values consisting of the stores we scraped, being Nikon's and Asics's brand stores and Amazon.com. This intercept also includes two cross-level interactions, between the marketplace level variable $Type\,of\,Marketplace_{jk}$ with the level 3 variables $Good\,Type_k$ and $Relative\,Price_k$. This level also has its own error term, e_{jk}, capturing all marketplace level variation that is unobserved. To complete our model, we formally define δ_{0k} in Equation 3 below.

$$\delta_{0k} = \beta_0 + \beta_1 * Good\,Type_k + \beta_2 * Relative\,Price_k + u_k \quad (3)$$

This intercept addresses the highest level variables in our model, the product k. These product levels are defined by the IDs we assigned to the same products in the data preparation process and consist of 34

different products. It has its own intercept, and two product level variables $GoodType_k$ and $RelativePrice_k$. This level also has its own error term, u_k, which captures unobserved differences in review valence between different products. By combining Equation 1, 2 and 3 we have our final model, which we can run to obtain our results.

Other than presenting a way to deal appropriately with our data structure, a multilevel model also allows us to control for unobserved product level differences. This is because, as explained above, the highest level intercept has an error term u_k that captures all variation in review valence between products that is unobserved. This is necessary, as the valence of reviews on a certain product is affected by more variables such as product quality for example, but we cannot capture this in our data.

5.2 Dealing with the dependent variable

We operationalize the dependent variable capturing a review's valence through two variables. The variable *Sentiment Score* is very straightforward, as it is a continuous score ranging from -1.8 to 7,8. *Star Rating* is a discrete variable as it is a rating on a scale of 1 to 5, with possible values of 1, 2, 3, 4 or 5. As such this requires a different model. We start with *Sentiment Score*, because this requires the simplest model out of the two.

Since this is a continuous score, we can model this relationship with a linear multilevel regression model (Schweidel & Moe, 2014). This means that the coefficients in our multilevel model above are approximated through a linear model. There are two important things to note here. First, we standardize the continuous independent variables as this is required for all multilevel models. Second, it must be noted that that multilevel models use maximum likelihood methods to estimate the coefficients and errors, which is different to the least squares approach OLS uses.

For the dependent variable *Star Rating* we must use a different model, as it is not a continuous variable. Earlier work by Y. Yang, Mao, and Tang (2018) and Binder, Heinrich, Klier, Obermeier, and Schiller (2019) addresses how to deal with star ratings specifically, so we can use this as a basis. While Binder et al. (2019) focuses more on the methodology, Y. Yang et al. (2018) shows an actual application of this method, trying to explain hotel review ratings through review- and hotel-level features. Another advantage of this paper is that these variables are also on different levels, requiring them to also use multilevel models. Our model below is thus largely based on their methodology. As we have a dependent variable here that is not only discrete but also ordered (since a rating of $1 < 2 < 3 < 4 < 5$), we use an ordered logistic regression model here. Like a regular logit model, this method takes the outcome of a linear model as input, which is plugged into a link function to arrive at a probability that an observation has a certain discrete value for the dependent variable. The output of this linear model is in our case y^*_{ijk}, and is obtained using the multilevel model described above. We generalize this model as follows:

$$y^*_{ijk} = \boldsymbol{\pi} * \boldsymbol{r_{ijk}} + \boldsymbol{\delta} * \boldsymbol{s_{jk}} + \boldsymbol{\beta} * \boldsymbol{x_k} + v_{ijk} + e_{jk} + u_k \tag{4}$$

Where $\boldsymbol{\pi} * \boldsymbol{x_{ijk}}$ contains all the review level variables and their coefficients, $\boldsymbol{\delta} * \boldsymbol{x_{jk}}$ contains all the store level variables and their coefficients and $\boldsymbol{\beta} * \boldsymbol{x_k}$ contains all the product level variables and their coefficients,

followed by each level's error term. To arrive at a probability y^* is plugged into the logistic link function to arrive at a probability as follows:

$$Pr = \frac{e^{y^*}}{1 + e^{y^*}} \tag{5}$$

Using these two equations, we build the ordered logit model shown in Equation 6.

$$\begin{aligned} Pr(y^*_{ijk} = l) &= Pr(t_{l-1} < y^*_{ijk} < t_l | \boldsymbol{r_{ijk}}, \boldsymbol{s_{jk}}, \boldsymbol{x_k}, v_{ijk}, ejk, u_k) \\ &= F(t_l - \boldsymbol{\pi} * \boldsymbol{r_{ijk}} - \boldsymbol{\delta} * \boldsymbol{s_{jk}} - \boldsymbol{\beta} * \boldsymbol{x_k} - v_{ijk} - e_{jk} - u_k) - \\ &\quad F(t_{l-1} - \boldsymbol{\pi} * \boldsymbol{r_{ijk}} - \boldsymbol{\delta} * \boldsymbol{s_{jk}} - \boldsymbol{\beta} * \boldsymbol{x_k} - v_{ijk} - e_{jk} - u_k) \end{aligned} \tag{6}$$

As this is an *ordered* logit, this equation models $Pr(y^*_{ijk} = l)$, the probability that a review with value y^*_{ijk}, has a star rating equal to l, where l takes values 1, 2, 3, 4 or 5. This is actually the probability that y^*_{ijk} falls between the two cutpoints t_{l-1} and t_l. For example, for a rating of $l = 4$ we want to know the probability that y^*_{ijk} has a value between the cutpoint of a rating of 3 and 4. The value of y^*_{ijk} is calculated using the multilevel model described above, and thus depends on the independent variables at the review level $\boldsymbol{r_{ijk}}$, those at the marketplace level, $\boldsymbol{s_{jk}}$ and those at the product level $\boldsymbol{x_k}$, and their respective residuals v_{ijk}, e_{jk} and u_k. To calculate this probability, this linear combination is first subtracted from each of the two cutoff points and plugged into the logistic link function F. This gives a value for each cutoff point, which are then subtracted from each other to get the probability that a review with value y^*_{ijk} gets a certain star rating. The coefficients of each variable in our output will tell us how that specific variable affects y^*_{ijk}, and thus the probabilities to fall within a star rating's range. This model, like the regular multilevel regression we run for *Sentiment Score*, requires continuous variables to be standardized and uses maximum likelihood to estimate these coefficients and cutpoints.

6 Results

In this section we present the results of this paper. We start with models using sentiment score as the dependent variable, followed by models using star rating as the dependent variable. For each we summarize the regression results, assess the hypotheses and perform several checks on assumptions and robustness.

6.1 Sentiment score as dependent variable

6.1.1 Estimation results

Following recommendations by Aguinis et al. (2013) we build up the model in steps. This means that we start with a null model that has no predictors and only consists of the residuals of each level. The goal of this null model is to see how much of the variation in sentiment is captured by each level. Table 5 below shows the variance of each level's residual term.

Level	Variance
Level 3: Product (k)	0.011
Level 2: Store (j)	0.002
Level 1: Review (i)	0.991

Table 5: Variance of each level in the null model

To get see how much of the variation in the sentiment score is captured by each level, we compute the IntraClass Correlation (ICC). This is computed by dividing a level's variance over the sum of all variances. In this case, no computation is needed as the variances sum to nearly 1 (1.004). As a result we can easily infer the ICCs from Table 5. The highest level, the product level, captures around 1% of total variation in the dependent variable, while the store level only captures 0.2% of the total variation. Both these scores are rather low, as most studies using only two-level multilevel models report ICC scores of around 15% for the higher level (Aguinis et al., 2013). This begs the question if using a multilevel model is really necessary here. We think it is, not only because are data structure is nested by nature but also because it allows us to better control for unobserved differences than an OLS regression would.

After the null model, we can now work towards the full model to assess our hypotheses. To do so, we first feed in the review level variables shown in Equation 1 to get our first model, while leaving out the cross-level interaction terms. For the second model, we add the store level variable *Type of Marketplace* to Model 1, while also leaving out the cross-level interaction terms. Third, we add the product level variables to get Model 3, after which we add all interaction terms at the product level to arrive at Model 4. Finally Model 5 also includes the review level interaction terms but not the product level interaction terms, as these are insignificant in Model 4. Table 6 below summarizes these models. For our results we let Model 2 and Model 5 be the focal models, as these are the models that include significant variables that assess the hypotheses in our conceptual framework.

Variable	Model 1	Model 2	Model 3	Model 4	Model 5
$Intercept$	0.033	0.081	0.035	0.006	-0.295
	(0.037)	(0.025)**	(0.045)	(0.051)	(0.106)**
$Dayssince_{ijk}$	-0.019	-0.023	-0.022	-0.019	0.040
	(0.009)*	(0.009)*	(0.009)*	(0.009)*	(0.017)*
$Number\ of\ Preceding\ Reviews_{ijk}$	-0.013	-0.003	0.002	-0.009	-0.562
	(0.014)	(0.013)	(0.014)	(0.018)	(0.176)**
$Review\ Helpfulness_{ijk}$	-0.044	-0.044	-0.044	-0.044	-0.044
	(0.007)***	(0.007)***	(0.007)***	(0.007)***	(0.007)***
$Review\ Length_{ijk}$	0.048	0.047	0.048	0.049	0.048
	(0.008)***	(0.008)***	(0.008)***	(0.008)***	(0.008)***
$Type\ of\ Marketplace_{ijk}$		-0.103	-0.105	-0.047	0.167
		(0.020)***	(0.021)***	(0.057)	(0.086)′
$Good\ Type_k$			0.068	0.095	0.134
			(0.054)	(0.060)	(0.058)*
$Relative\ Price_k$			0.002	0.006	0.007
			(0.017)	(0.020)	(0.018)
$Type\ of\ Marketplace_{jk} * Good\ Type_k$				-0.064	
				(0.058)	
$Type\ of\ Marketplace_{jk} * Relative\ Price_k$				-0.007	
				(0.019)	
$Type\ of\ Marketplace_{jk} * Dayssince_{ijk}$					-0.085
					(0.020)***
$Type\ of\ Marketplace_{jk} * Number\ of\ Preceding\ Reviews_{ijk}$					0.600
					(0.181)***

Note: Standard errors are in parentheses
$***p < 0.001$, $**p < 0.01$, $*p < 0.05$, $'p < 0.1$

Table 6: Regression results of Models 1-5, using sentiment score as dependent variable

To assess our main effect, we first look at Model 2. $Type\ of\ Marketplace_{ijk}$ shows a negative coefficient of -0.103 that is significant at a 0.1% level. This implies that reviews in a third party online channel are on average more negative than those in a direct online channel, keeping all other variables constant. In our second focal model, Model 5, we now see a slightly positive coefficient of 0.167 that is significant at a 10% level. Its significance is thus questionable, but it implies that reviews in a third party online channel are on average more positive than those in a direct online channel, keeping all other variables constant. However, keeping all other variables constant is impossible here, as the interaction terms also come into play here when $Type\ of\ Marketplace_{ijk}$ changes. Therefore we assess the main effect using Model 2 which allows us to accept H1.

Having addressed the main effect, we can now assess the moderating effects. To assess the product level moderators $Good\ Type_k$ and $Relative\ Price_k$ we use Model 4 which shows coefficients for the relevant interaction terms that are not significant. We can conclude that at least for sentiment score there is no moderating effect here, meaning that we find no evidence for H2 and H3. For the review level moderators $Dayssince_{ijk}$ and $Number\ of\ Preceding\ Reviews_{ijk}$, Model 5 shows coefficient for both variables' interactions with $Type\ of\ Marketplace_{ijk}$ that are significant at a 0.1% level. For the interaction of $Dayssince_{ijk}$ with the main effect variable we see a negative coefficient of -0.085 which implies that in a third party online channel reviews are on average more negative than in a direct channel when looking at older reviews. However, this only holds when keeping all other variables constant, which is not the case here as the variables $Type\ of\ Marketplace_{ijk}$ and $Dayssince_{ijk}$ also change. The former has a coefficient of 0.167 that

is only significant at a 10% level, while the latter has a coefficient of 0.040. This indicates that equally aged reviews are on average more positive in the third party channel, while older reviews are on average more negative (positive) in the third party (direct channel). For older reviews this initial positive difference between the marketplaces thus decreases. This means that we reject H4a as this expects that there is a *negative* difference in early reviews that becomes smaller for older reviews. For the interaction term of $Number\ of\ Preceding\ Reviews_{ijk}$ there is a significant coefficient of 0.600, indicating that on average reviews in the third party channel are more positive when there are more reviews preceding them. Again, to fully understand this effect we must also take into account the coefficients of the standalone variables. For $Type\ of\ Marketplace_{ijk}$ this is again 0.167 ($p < 0.1$) and for $Number\ of\ Preceding\ Reviews_{ijk}$ this is equal to -0.562 ($p < 0.01$). This means that again for reviews with equal $Number\ of\ Preceding\ Reviews_{ijk}$ the third party channel is more positive, and these two diverge as there are more reviews. This looks more in line with H4b but we still reject it. Even though in the direct channel reviews become more negative as there are more, while in the third party channel the opposite takes place, we need to observe an initial negative difference to accept H4b. Interestingly these two moderators' effects oppose each other: when there are few reviews in one day on a product $Dayssince_{ijk}$ dominates and when there are many reviews on a product in one day $Number\ of\ Preceding\ Reviews_{ijk}$ dominates.

Finally, the review level control variables that show interesting coefficients in Models 1-5 are $Dayssince_{ijk}$, $Review\ Helpfulness_{ijk}$ and $Review\ Length_{ijk}$. The first has a negative coefficient that is significant at a 5% level in the models where it does not have an interaction term as well, which indicates that as reviews are relatively older, they tend to become more negative on average, keeping all other variables constant. $Review\ Helpfulness$ shows an even higher significance level at 0.1% and shows that on average reviews that are voted as more helpful tend to be more negative in valence. This is consistent with what we expected, and the reason we control for it. Not consistent with our expectations is the significant positive coefficient of $Review\ Length_{ijk}$. It shows that on average longer reviews are more positive in valence, while we expected the opposite (Homburg et al., 2015). Nevertheless, its significance shows that it is worth controlling for this.

6.1.2 Assumptions

Like every other statistical method multilevel linear regressions require certain assumptions to hold for its results to hold. These are the same assumptions required for OLS regressions (Winter, 2013). The first assumption is that of independence: the data points and variables in our dataset must be independent of each other. We already discussed this issue earlier in the methodology and use multilevel modeling exactly for this purpose, as our data is by definition not independent. As we adjusted our methodology to solve for this we can be confident that this assumption is not violated. Another assumption regarding the data structure addresses multicollinearity. Again, we already discussed this in the methodology section and designed our model in such a way that this should not be an issue. Also, we checked for variables with excessive correlation in the data section and found no issues there.

The next assumption is that of linearity: the relation between the dependent and independent variable must be linear, or we must specify the model differently. To check this assumption we use the residual plot below.

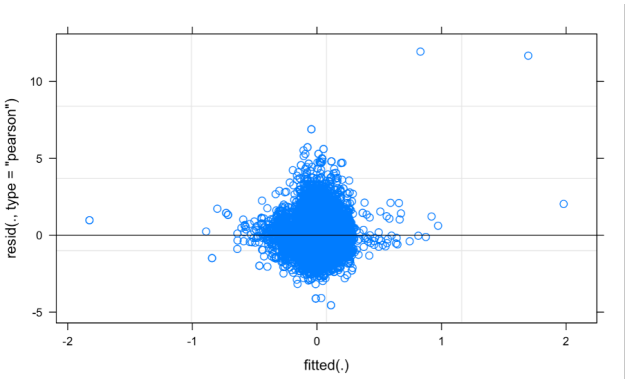

Figure 3: Residual plot of Model 5, using sentiment score as the dependent variable

If the assumption of linearity holds, there should be no clear pattern in this plot: all residuals should be in the same interval around 0. However, this is very hard to observe in Figure 3 because there are four large outliers. This is another assumption of linear multilevel models that we first check before continuing with the others. Outliers or influential data points should not be present, but they clearly are. Further investigation shows that the residual in the very left part of Figure 3 is a very negative review, that also had a 1 star rating with it. The observations in the top right received very high sentiment scores from the `polarity` algorithm and were also very long. However, they only had a 3 and 4 star rating respectively so it seems that the sentiment score was influenced by the very large amount of words indicating sentiment. For the residual on the right of Figure 3 the same holds as for the one on the left but opposite: it is very positive. In the next subsection we check our models' robustness and look how our results change if we remove these outliers and reestimate our models.

Continuing with the assumptions while ignoring these outliers, we get Figure 4 below. We now can assess the linearity assumption and indeed this looks a lot better: the residuals are approximately randomly spread through the plot. The only thing to notice is that there are fewer residuals for both lower and higher fitted values. The next assumption to check is the assumption of homoskedasticity: the residuals must be the same across all values of the independent variables. We check this in the same way as the linearity assumption, by looking at Figure 4 below. Here the residuals must be distributed randomly throughout the plot which they approximately are: there seems to be no clear pattern.

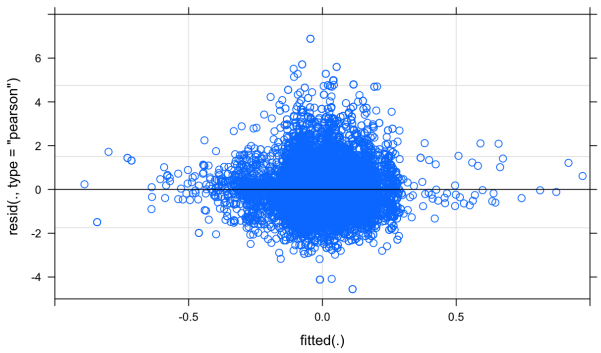

Figure 4: Residual plot of Model 5 on a different x-axis scale, using sentiment score as the dependent variable

Another assumption regarding the residuals is the normality assumption: we need the residuals to be distributed normally. To check this we use the Q-Q plot in Figure 5 below. Apart from the outliers that are the same points we observed before, this Q-Q plot looks good. All points follow an approximately straight, diagonal line throughout the plot.

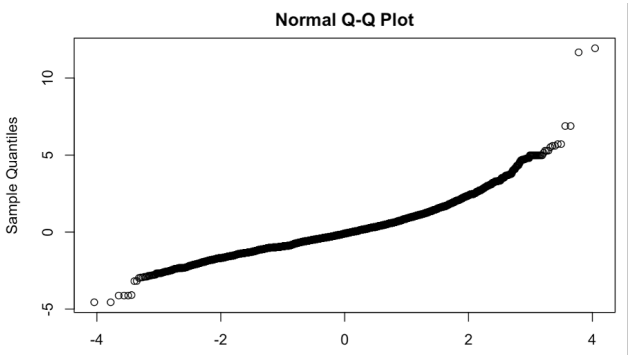

Figure 5: Q-Q plot of the residuals Model 5, using sentiment score as the dependent variable

6.1.3 Robustness tests

The first robustness check we perform concerns the outliers we found above. We remove these four observations from our data and reestimate Models 1-5. The results are in Table 7 below. Comparing this table with the original Table 6, there are only minor differences. The results that we need to assess our hypotheses are robust, and the only changes that occur in these variables is a slightly different estimates of the coefficients and standard errors. The only major change here occurs in $ReviewLength_{ijk}$, whose coefficients are no longer significant in all models. It makes sense that this variable changes, as the outliers were all very long reviews.

Variable	Model 1	Model 2	Model 3	Model 4	Model 5
$Intercept$	0.041	0.084	0.076	0.056	-0.244
	(0.036)	(0.024)**	(0.044)′	(0.050)	(0.104)*
$Dayssince_{ijk}$	-0.026	-0.028	-0.028	-0.026	0.036
	(0.009)**	(0.009)**	(0.009)**	(0.009)**	(0.017)*
$Number\ of\ Preceding\ Reviews_{ijk}$	-0.012	-0.005	-0.005	-0.013	-0.551
	(0.014)	(0.013)	(0.014)	(0.017)	(0.174)**
$Review\ Helpfulness_{ijk}$	-0.042	-0.042	-0.043	-0.043	-0.043
	(0.008)***	(0.008)***	(0.008)***	(0.008)***	(0.008)***
$Review\ Length_{ijk}$	-0.008	-0.009	-0.008	-0.008	-0.009
	(0.008)	(0.009)	(0.009)	(0.009)	(0.009)
$Type\ of\ Marketplace_{ijk}$		-0.103	-0.103	-0.061	0.160
		(0.020)***	(0.020)***	(0.057)	(0.085)′
$Good\ Type_k$			0.012	0.031	0.075
			(0.053)	(0.058)	(0.056)
$Relative\ Price_k$			0.000	0.005	0.005
			(0.017)	(0.020)	(0.017)
$Type\ of\ Marketplace_{jk} * Good\ Type_k$				-0.045	
				(0.057)	
$Type\ of\ Marketplace_{jk} * Relative\ Price_k$				-0.010	
				(0.019)	
$Type\ of\ Marketplace_{jk} * Dayssince_{ijk}$					-0.088
					(0.020)***
$Type\ of\ Marketplace_{jk} * Number\ of\ Preceding\ Reviews_{ijk}$					0.582
					(0.179)**

Note: Standard errors are in parentheses
$***p < 0.001$, $**p < 0.01$, $*p < 0.05$, $'p < 0.1$

Table 7: Regression results of Models 1-5 excluding outliers, using sentiment score as dependent variable

We can thus stay with Table 6 as the output of our analysis that uses sentiment as the dependent variable, but there are some final robustness tests we can perform. As with all regressions we use statistical significance of coefficients as a basis to interpret the effect or not. However, statistical significance of coefficients and thus p-values are not straightforward when it comes to multilevel models. To build our models we use the `lme4` package by Bates, Mächler, Bolker, and Walker (2014), who have made what they call the controversial decision to not include p-values in their package. This is because the null distributions of these multilevel models are not t-distributed, while it also hard to estimate the degrees of freedom one would need. To obtain the p-values we discussed above we use an additional package called `lmerTest`, which does return p-values using t-testing where the degrees of freedom are estimated using Satterthwaite approximation (Kuznetsova, Brockhoff, & Christensen, 2017). This technique is referred to as an ad hoc solution by Bates et al. (2014), so it may be worthwhile to look a bit more in depth at how we obtain significance of coefficient, to really establish our results.

As this issue of p-values is relevant for many researchers who wish to use multilevel models to make inferences, Luke (2017) outlines four techniques one could do this. These methods are likelihood ratio tests, using the Wald t-statistics that lme4 does report as z-statistics (z-as-t), Satterthwaite approximation (which we apply above) and parametric bootstrapping. We consider the latter outside of the scope of this paper, but go ahead with likelihood ratio test and z-as-t, to see how robust our estimates' significance are. We do this for $Type\ of\ Marketplace_{jk}$ in Model 3 and its interactions with $Dayssince_{ijk}$ and $Number\ of\ Preceding\ Reviews_{ijk}$ in Model 5.

Starting with the likelihood ratio test, the idea here is to build two models: one without and one with the variable whose coefficient we want to establish the significance of. We do this using the `anova` function, which tests the hypothesis that adding that variable adds no explanatory power using a χ^2 distribution. It returns a p-value, which allows us to conclude that including that variable adds significant explanatory power when it is lower than 0.05. Table 8 below shows these p-values, where Model 3 indicates a test of Model 3 without $Type of Marketplace_{jk}$, Model 5a indicates Model 5 without $Type of Marketplace_{jk} * Dayssince_{ijk}$ and Model 5b indicates Model 5 without $Type\ of\ Marketplace_{jk} * Number\ of\ Preceding\ Reviews_{ijk}$.

Variable	Model 3	Model 5a	Model 5b
$Type\ of\ Marketplace_{jk}$	0.014*		
$Type\ of\ Marketplace_{jk} * Dayssince_{ijk}$		0.000***	
$Type\ of\ Marketplace_{jk} * Number\ of$ $Preceding\ Reviews_{ijk}$			0.001***

Note: $***p < 001, **p < 0.01, *p < 0.05$

Table 8: P-values of each of likelihood ratio tests

Using this likelihood ratio test, it seems that our results hold as all p-values are below 0.05. Secondly, we consider using the t-statistics as z-statistics to get p-values for the coefficients. Again, we do this for the same variables as in Table 8. Though we did not show them above in the estimation results, the `lme4` did report t-statistics, which are shown for the variables and their respective models we investigate in Table 9.

Variable	t-statistics, used as z-statistic	p-value
$Type\ of\ Marketplace_{jk}$ (In Model 3)	-5.129	0.000
$Type\ of\ Marketplace_{jk} * Dayssince_{ijk}$ (In Model 5)	-4.184	0.000
$Type\ of\ Marketplace_{jk} * Number\ of\ Preceding$ $Reviews_{ijk}$ (In Model 5)	3.317	0.000

Table 9: Results of likelihood ratio tests

Again we can easily that these coefficients are significant using the t-as-z approach. We are thus confident in our results and estimates, but should be noted that p-values in the context of multilevel models are still developing at the time of writing (Bates et al., 2014; Luke, 2017).

6.2 Star Rating as dependent variable

For the second and last part of the results section, we assess the same effects we outlined in our framework but now use star rating as the dependent variable. As discussed in the methodology section, we use an ordered logistic regression to do this. In R this means that we fit a so-called cumulative link mixed model using the `ordinal` package (Christensen, 2019a). The results of this model are similar to the linear multilevel model, but the interpretation differs. In this case a coefficient indicates how a change in that variable affects the input of the logistic link function, which models the probability that a review has a certain star rating. As we only care about testing the effects outlined in our framework, we need not worry about this and only

look for significant positive or negative coefficients. In building the model we follow recommendations in a tutorial accompanying this package by its author Christensen (2019b). This states that we need not run a null model first, so we build the model straightaway using the same approach we did for sentiment score. The results are show in Table 10 below. The focal models we discuss here are Model 3, 4 and 5 as these all include significant coefficients that help in assessing our hypotheses.

Variable	Model 1	Model 2	Model 3	Model 4	Model 5
$Dayssince_{ijk}$	-0.043	-0.043	-0.035	-0.037	0.106
	(0.024)′	(0.024)′	(0.024)	(0.024)	(0.046)*
$Number\ of\ Preceding\ Reviews_{ijk}$	0.035	0.037	-0.057	-0.048	-1.646
	(0.045)	(0.045)	(0.057)	(0.047)	(0.424)***
$Review\ Helpfulness_{ijk}$	-0.055	-0.055	-0.056	-0.055	-0.056
	(0.016)***	(0.016)***	(0.016)***	(0.016)***	(0.016)***
$Review\ Length_{ijk}$	-0.219	-0.219	-0.226	-0.228	-0.227
	(0.020)***	(0.020)***	(0.020)***	(0.020)***	(0.020)***
$Type\ of\ Marketplace_{ijk}$		-0.192	-0.040	-0.270	0.635
		(0.342)	(0.211)	(0.162)′	(0.283)*
$Good\ Type_k$			-0.806	-1.153	-0.935
			(0.195)***	(0.171)***	(0.186)***
$Relative\ Price_k$			-0.058	0.014	0.026
			(0.051)	(0.060)	(0.062)
$Type\ of\ Marketplace_{jk} * Good\ Type_k$				0.399	0.229
				(0.161)*	(0.167)
$Type\ of\ Marketplace_{jk} * Relative\ Price_k$				-0.109	-0.116
				(0.052)*	(0.052)*
$Type\ of\ Marketplace_{jk} * Dayssince_{ijk}$					-0.170
					(0.053)**
$Type\ of\ Marketplace_{jk} * Number\ of\ Preceding\ Reviews_{ijk}$					1.647
					(0.429)***

Note: Standard errors are in parentheses
$***p < 0.001$, $**p < 0.01$, $*p < 0.05$, $'p < 0.1$

Table 10: Regression results of Models 1-5, using star rating as dependent variable

Starting with our main effect through $Type\ of\ Marketplace_{ijk}$ we see that the coefficient is insignificant for all models except Models 4 ($p < 0.1$) and 5 ($p < 0.05$). However, this could be caused by the fact that it correlates strongly with the four interaction terms in that same equation, and it is not enough for us to accept H1.

Moving to the moderators, we first see significant coefficients for the product level interaction effects in Model 4. $Type\ of\ Marketplace_{jk} * Good\ Type_k$ reports a coefficient of 0.399, which implies that for experience goods the star rating in the third party channel is higher than its is in the direct channel. To fully understand the effect we start with $Type\ of\ Marketplace_{jk}$ that has a coefficient of -0.270 significant at a 10% level, meaning that a review on a search good (for which $Good\ Type_k$ is 0) has on average a lower star rating if it is posted in a third party channel. $Good\ Type_k$ has a significant coefficient of -1.153, meaning that a review on an experience good in a direct channel (when $Type\ of\ Marketplace_{jk}$ is 0) has a lower star rating than a search good in that same channel. Now returning to the moderating effect captured by $Type\ of\ Marketplace_{jk} * Good\ Type_k$, its significant coefficient of 0.399 indicates that for an experience good the star rating actually becomes higher in the third party channel compared to the direct channel. This

means that we reject H2, as this expected the star rating of a review on an experience good in the third party channel to be even lower relative to that in the direct channel. For $Type\,of\,Marketplace_{jk} * Relative\,Price_k$ we see a significant coefficient of -0.109, indicating that when a product is more expensive in a third party channel compared to the direct channel, the star review is lower on average in that third party channel. This means that we accept H3. Interestingly, the coefficient of $Relative\,Price_k$ is insignificant, which may imply that consumers in the direct channel let their reviews not be affected by any price difference that may exist.

For the review level moderators in Model 5 we see significant coefficients that oppose each other, just like we did with sentiment score as the dependent variable. $Type\,of\,Marketplace_{jk} * Dayssince_{ijk}$ has a coefficient of -0.116, which implies that for older reviews the star rating is lower in the third party channel compared to the direct channel. Again this holds only when keeping all other variables constant, which is not the case as $Type\,of\,Marketplace_{jk}$ and $Dayssince_{ijk}$ also change and have significant coefficients in Model 5. $Type\,of\,Marketplace_{jk}$ has a significant coefficient of 0.635, implying that on average a review's star rating is higher in the third party channel for relatively new reviews. Moreover, $Dayssince_{ijk}$ has a significant coefficient of 0.106 meaning that older reviews on average are more positive in the direct channel. Then, the significant coefficient of $Type\,of\,Marketplace_{jk} * Dayssince_{ijk}$ of -0.116 indicates that older reviews have lower star ratings in the third party channel. This implies that the initial positive difference between a third party channel and direct channel review is smaller or even negative with older reviews. This completely opposes H4a, which expects a negative difference that then disappears with older reviews. We therefore reject H4a. For $Type\,of\,Marketplace_{jk} * Number\,of\,Preceding\,Reviews_{ijk}$ we see a coefficient of 1.647. $Number\,of\,Preceding\,Reviews_{ijk}$ has a significant coefficient of -1.646, meaning that in case of a third party channel these effects cancel out and valence stays constant. Therefore, when there are more reviews present in the direct channel we expect to see more negative reviews while in the third party channel this does not take place. However, for us to accept H4b we would need the valence to be more negative in the third party channel for the earlier reviews, which they are not. We therefore reject H4b.

Finally for the review level control variables we see that similar to when we had sentiment score as dependent variable, $Review\,Helpfulness_{ijk}$ and $Review\,Length_{ijk}$ have coefficients that are significant. For review helpfulness the direction is also the same as shown by a coefficient of -0.055/-0.056, indicating that reviews with more votes tend to have lower star ratings on average, keeping other variables constant. For $Review\,Length_{ijk}$ the coefficient is opposite with what we saw with sentiment score, as is now a negative coefficient round -0.2. However, this variables estimate's are likely heavily influenced by the outliers in our data we observed in the previous model's assumption checks, so it is not worthwhile to interpret it.

6.2.1 Assumptions

A multilevel logistic regression model also has its assumptions that we check in this final section of the results. The first assumption is concerns the dependent variable as it requires a dependent variable that is indeed ordered. In this case this is clearly satisfied, as the star rating moves from 1 (very negative) to 5 (very positive). Moreover, the independence and multicollinearity assumptions must be satisfied here. As discussed above, we already adjusted our data and model in such a way that these assumptions should hold.

The linearity, homoskedasticity and normality assumptions that all address the residuals are not required

for logistic models in general. However, ordered logistic models have an additional assumption that also applies to our models. This is the proportional odds assumption, which requires the relationship between the dependent variable and the independent variables (i.e. the coefficients) to be the same for all levels of the dependent variable (i.e. each star rating). The problem here is that there is currently no method available to check this assumption for our method. We fit a three level cumulative link mixed model using the `clmm` function in the `ordinal` package, which only includes ways to test this assumption for two levels. The `clmm` function is rather new so testing this assumptions is something that will hopefully be possible in the future (Christensen, 2019a).

6.2.2 Robustness tests

As with the linear model above, this multilevel ordered logit model has p-values that are not necessarily straightforward. The `ordinal` package does return p-values by default, which we used in Table 10. These are based on z-statistics, as is common practice with logistic regression (Bates et al., 2014; Christensen, 2019a). Another way to assess significance is again a likelihood ratio test, as this is highly recommended by Christensen (2019b). Just like with sentiment score as the dependent variable, we use the `anova` function to compute the relevant p-values in Table 11. However, we now not only do it for $Type\,of\,Marketplace_{jk}$, $Type\,of\,Marketplace_{jk} * Dayssince_{ijk}$ and $Type\,of\,Marketplace_{jk} * Number\,of\,Preceding\,Reviews_{ijk}$ but also for $Type\,of\,Marketplace_{jk} * Good\,Type_k$ and $Type\,of\,Marketplace_{jk} * Relative\,Price_k$. This means that we rerun Model 3 for $Type\,of\,Marketplace_{jk}$ and Model 4 and Model 5 for the product and review level moderators respectively. Consistent with what we already saw, these tests show that $Type\,of\,Marketplace_{jk}$ on its own has no significant effect, and all moderating variables do.

Variable	Model 3	Model 4a	Model 4b	Model 5a	Model 5b
$Type\,of\,Marketplace_{jk}$	0.844				
$Type\,of\,Marketplace_{jk} * Good\,Type_k$		0.013*			
$Type\,of\,Marketplace_{jk} * Relative\,Price_k$			0.037*		
$Type\,of\,Marketplace_{jk} * Dayssince_{ijk}$				0.001***	
$Type\,of\,Marketplace_{jk} * Number\,of\,Preceding\,Reviews_{ijk}$					0.000***

Note: $***p < 0.001$, $**p < 0.01$, $*p < 0.05$, $'p < 0.1$

Table 11: P-values of each of likelihood ratio tests

7 Conclusion and discussion

In this paper we set out to uncover the differences in online product reviews over marketplaces that are owned by a manufacturer and marketplaces that are not owned by a manufacturer and sell many other brands as well. In addition, we attempted to show how this difference is moderated depending on the type of good, price differences between the marketplaces and the relative timing and position of a review. In this part of the paper we summarize our findings and put them into perspective. Finally, we discuss this paper's limitations and avenues for future research. To provide an overview of our findings we summarize them in Table 12 below.

Main effect	Hypothesis	Finding
Type of Marketplace	H1: The valence of reviews on the same product is more positive in direct online sales channels compared to third party online sales channels	Accepted for sentiment score
Moderating effects	**Hypothesis**	**Finding**
Good Type: Search or Experience	H2: The difference in review valence for the same product between a direct online sales channel and a third party online sales channel is larger for reviews about experience goods compared to reviews about search goods.	Rejected
Relative Price	H3: The difference in review valence for the same product between a direct online sales channel and a third party online sales channel is smaller (larger) when the price of that product is higher (lower) in the direct online sales channel, compared to the third party online sales channel.	Accepted for star rating
Relative timing of a review	H4a: The difference in review valence for the same product between a direct online sales channel and a third party online sales channel is larger for reviews written earlier in time compared to reviews written later in time	Rejected
Relative position of a review	H4b: The difference in review valence for the same product between a direct online sales channel and a third party online sales channel is larger for reviews that have few reviews preceding them compared to reviews that have many reviews preceding them	Rejected

Table 12: Summary of this paper's findings

Starting with the main effect, we found evidence that supports H1 by showing that in the manufacturer owned, direct channel review *sentiment* is indeed more positive than in the third party channel. Note that we say sentiment score, as we did not such evidence for star rating. Nevertheless this is interesting for managers to know, as Hu et al. (2014) show that reviews can also affect sales through their underlying sentiment. For the star ratings we found very mixed evidence, as the type of marketplace was insignificant for some models but changed signs when interactions where introduced. We saw this for sentiment score as well, and this can be attributed to the fact that these interactions correlate strongly with the marketplace type. It shows that the difference between marketplaces depends strongly on the circumstances, which we looked at through the moderating effects.

Regarding the moderators, we found little evidence that supported our hypotheses. Starting with the product level moderator of experience versus search goods we found only for star rating that experience goods tend to receive lower ratings than search good in the direct channel, and that this difference is smaller for third party channels. This is not what our hypothesis expected, as we expected the self-selection bias of positive customers in the direct channel to be stronger for experience goods and thus even more positive reviews for experience goods compared to the third party channel. Why is then valence a lot more negative for experience goods in the direct channel? Going back to the meaning of an experience good, it is a good

whose quality can only be fully assessed after purchase, based on subjective factors that must be experienced personally (Cui et al., 2012; Mudambi & Schuff, 2010). As a result consumers cannot fully evaluate the product based on product information and available product reviews only. Combined with a potential self-selection bias that leads to overly positive reviews in especially the direct channel, this can then lead to customers buying product with high expectations, only to be disappointed after purchase. According to motivation theory, they are then motivated to signal their contrasting opinion and then write a more negative review on that experience good (Wu & Huberman, 2008). This mechanism can thus lead to more negative reviews in the direct channel for that experience good, and it may be less strong in the third party channel reviews. Regarding the moderator of relative price, we found evidence regarding the intuitive H3: when a product is relatively more expensive in the third party channel, customers write relatively more negative reviews in that marketplace because their utility is decreased by that higher price (Li & Hitt, 2010).

For the the review level moderators we reject H4a: for older reviews we see the difference in valence between the marketplaces become larger, with older reviews in the third party channel being more negative. We expected the opposite, as we expected the reviews to become more negative in the direct channel and valence in both channels to approach each other. We did see such a trend for sentiment score and the sheer number of reviews on a product, hinting at H4b. Review valence between the two marketplaces does approach each other here as there are more reviews, but we do not observe the initial difference we expected. Also, it is hard to fully assess whether the mechanism we expected in our framework takes place here, as this is conditional on the circumstances. For a product that receives many reviews in one day, the number of reviews will increase more than the number of days variable, so in that case the trend we expect takes place. For popular products our framework seems to be confirmed, but for unpopular products it does not.

It is important to recognize the limitations of this study, as others might be able to improve it. The first limitations concern the data used. The design of the study made it a very labor intensive task to collect data. As a result we only had two brands for two marketplaces and two good types, while a study that has many brands and marketplaces may be able to draw conclusions that are more externally valid. Also, matching reviews on the same products across marketplaces proved to be difficult. Amazon uses a uniform, easy-to-use system of *ASIN* codes to classify its products, but these were obviously not used in the other marketplaces. For Nikon matching could be checked by hand, as there were not many unique products. Especially for Asics there could be room for a more waterproof system, as Asics also uses its own product codes. We attempted to use these for the matching process, but Amazon only sparsely or incompletely reported these.

Moreover, there are some limitations regarding the methodology used. We used multilevel models to appropriately deal with our data structure and control for unobserved differences in valence across different products. More specifically we use random intercept models, where the intercepts have random effects and the independent variables have fixed effects (see Snijders (2005) and Snijders and Bosker (2011) for a complete explanation of fixed vs. random effects). This means that in our models the relationship between the predictors and valence is assumed to be constant across reviews (regardless of their products and stores), while only the intercept varies based on a review's product and store. As a result we ignore any heterogeneity that may exist in the relationship between our predictors and valence. This is a limitation to our study that must be recognized, and something that future work could expand upon. Finally for the multilevel ordered logit, or cumulative logistic mixed model, we used an advanced method for which not all necessary

functionalities are yet available. This means that we were not able to check its proportional odds assumption.

As a final part of this paper it is interesting to discuss what avenues for future research this paper provides. We set out to show if and how review valence differs across two types of marketplaces, and found that in term of sentiment score it does. However, we found little evidence for the moderating effects in our conceptual framework, meaning that the 'how' part of our attempt is still very open. This could be interesting for future endeavours: we showed managers that the difference is there, but to fully help in the sales channel decision it is good to know exactly why it is happening. Also, we think that any research attempt that can appropriately deal with the limitations above can enrich to our findings. It is for instance interesting to see how these finding hold when looking at an entire industry of brands, or the entire portfolio of a brand's sales channels. Finally, we think that topic modeling can be of help in answering the why part of our questions. As shown in the appendix, we attempted to show how topics discussed in reviews differ across marketplaces in hopes of discovering where these differences lie and why. However, our attempt returned inconclusive results, so a work that focuses entirely on discovering relevant topics and contrasting them across marketplace types could add a lot of value. In sum, we wanted to introduce a new angle to look at the sales channel choice by looking at customer reviews, and now that this has been done future work can expand our understanding of it.

8 Appendix

Posterior study: Topic modeling to uncover differences

In this paper we managed to show that there are indeed differences in valence of online product reviews between a direct and third party channel, but we were unable to show exactly why that is the case. In this appendix section we show another attempt that was made outside of the scope of the original framework and methodology to give our conclusion more depth. This attempt consists of topic modeling: uncovering latent topics in the our body of reviews and computing how much each review is about that topic. These results can then be used to run another regression where we model to what extent certain topics are likely appear in a review in the third party marketplace versus the direct channel. We summarize the results of this effort in this appendix section, but did not include it in the man paper because its findings turn out to be uninsightful.

To model these topics we use Latent Direchlet Allocation (LDA). LDA is called a *generative* method, as it tries to model how texts are generated. It assumes that when writing a review, a writer randomly chooses a topic using a prespecified distribution of topics for that review. Then, to get the words for that review the writer randomly chooses a word randomly using the distribution of words for that topic (Blei, 2012). This distribution of reviews over topics and topics over words is what LDA aims to model, it tries to find out the hidden structure behind reviews. It does this based on the words it observes in the reviews, meaning that it actually reverses this generative process. More specifically, LDA models review i's distribution over k topics as a vector of Dirichlet probabilities θ_i and each topic's distribution over words as a vector of Dirichlet probabilities β_k.

We seperate the reviews from Nikon and Asics, as otherwise topic differences may be driven by these product differences. Then, we let LDA model 15 topics, as this has been done before by several other studies in the same field (Liu et al., 2017). For each review we then have 15 new variables, each indicating the the probability that that review belongs to that topic. We then feed this into a logistic regression where *Type of Marketplace* is the dependent variable and the topic probabilities are the independent variables. The output of this regression is shown below for Nikon cameras.

Variable	Estimate	Standard error
Intercept	-0.246	0.223
Topic 1	2.221	0.364***
Topic 2	4.074	0.500***
Topic 3	1.172	0.584*
Topic 4	2.010	0.303***
Topic 5	1.623	0.403***
Topic 6	0.109	0.380
Topic 7	0.923	0.375*
Topic 8	0.159	0.370
Topic 9	1.698	0.373***
Topic 10	-0.035	0.275
Topic 11	1.109	0.360**
Topic 12	1.078	0.304***
Topic 13	3.443	0.505***
Topic 14	-0.033	0.384

Note: $***p < 0.001$, $**p < 0.01$, $*p < 0.05$, $'p < 0.1$

Table 13: Results of our topic modeling analysis for Nikon

Note that Topic 15 is excluded here to avoid multicollinearity, and functions as the reference category here. What this analysis broadly tells us, is that Topics 1, 2, 4, 5, 9, 11, 12 and 13 are topics that make a review a lot more likely to be in the third party channel. To see what these topics mean we look at the top 10 words associated with each, shown in Figure 6 below. These latent topics can be interpreted by the researcher, as for instance Topic 1 can refer to usability, and topic 4 can refer to the product's quality. However, between these topics there is no clear division over interpretable topics, so this attempt of topic modeling falls short. It is, however, an interesting avenue for future research to explore as topic modeling has shown its value in earlier marketing literature (Reisenbichler & Reutterer, 2019).

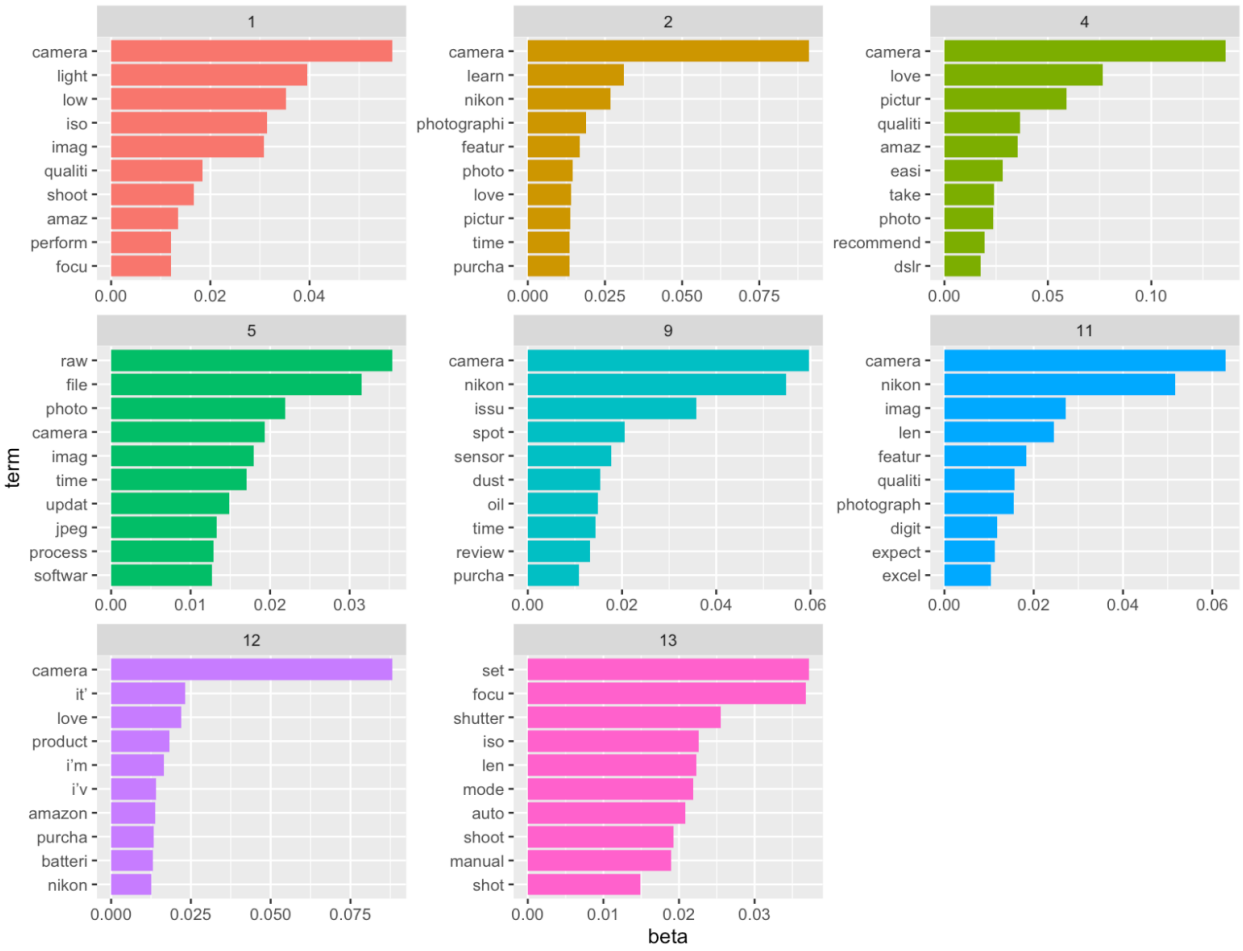

Figure 6: Top words for topics 1, 2, 4, 5, 9, 11, 12 and 13

9 References

Aaker, D. A. (2009). *Managing brand equity.* Simon and Schuster.

Aguinis, H., Gottfredson, R. K., & Culpepper, S. A. (2013). Best-practice recommendations for estimating cross-level interaction effects using multilevel modeling. *Journal of Management*, *39*(6), 1490–1528.

Arya, A., Mittendorf, B., & Sappington, D. E. (2007). The bright side of supplier encroachment. *Marketing Science*, *26*(5), 651–659.

Bambauer-Sachse, S., & Mangold, S. (2011). Brand equity dilution through negative online word-of-mouth communication. *Journal of Retailing and Consumer Services*, *18*(1), 38–45.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*.

Beneke, J., de Sousa, S., Mbuyu, M., & Wickham, B. (2016). The effect of negative online customer reviews on brand equity and purchase intention of consumer electronics in south africa. *The International Review of Retail, Distribution and Consumer Research*, *26*(2), 171–201.

Bhavaraju, S. K. T., Beyney, C., & Nicholson, C. (2019). Quantitative analysis of social media sensitivity to natural disasters. *International Journal of Disaster Risk Reduction*, *39*, 101251.

Binder, M., Heinrich, B., Klier, M., Obermeier, A. A., & Schiller, A. (2019). Explaining the stars: Aspect-based sentiment analysis of online customer reviews.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77–84.

Bruhn, M., Schoenmueller, V., & Schäfer, D. B. (2012). Are social media replacing traditional media in terms of brand equity creation? *Management Research Review*, *35*(9), 770–790.

Cai, G. G. (2010). Channel selection and coordination in dual-channel supply chains. *Journal of Retailing*, *86*(1), 22–36.

Cheah, B. C. (2009). Clustering standard errors or modeling multilevel data. *University of Columbia*, 2–4.

Chen, J. (2019). Creating an online review management strategy. *Sprout Social*. Retrieved 2021-04-14, from `https://sproutsocial.com/insights/online-review-management/`

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

Chiang, W.-y. K., Chhajed, D., & Hess, J. D. (2003). Direct marketing, indirect profits: A strategic analysis of dual-channel supply-chain design. *Management Science*, *49*(1), 1–20.

Christensen, R. H. B. (2019a). *ordinal—regression models for ordinal data.* (R package version 2019.12-10. https://CRAN.R-project.org/package=ordinal)

Christensen, R. H. B. (2019b). *A tutorial on fitting cumulative link mixed models with clmm2 from the ordinal package.* Retrieved from `https://cran.r-project.org/web/packages/ordinal/vignettes/clmm2_tutorial.pdf`

Cui, G., Lui, H.-K., & Guo, X. (2012). The effect of online consumer reviews on new product sales. *International Journal of Electronic Commerce*, *17*(1), 39–58.

Dellarocas, C. (2006). Strategic manipulation of internet opinion forums: Implications for consumers and firms. *Management Science*, *52*(10), 1577–1593.

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.

Eelen, J., Özturan, P., & Verlegh, P. W. (2017). The differential impact of brand loyalty on traditional and online word of mouth: The moderating roles of self-brand connection and the desire to help the brand. *International Journal of Research in Marketing*, *34*(4), 872–891.

Fernandes, E., Moro, S., Cortez, P., Batista, F., & Ribeiro, R. (2021). A data-driven approach to measure restaurant performance by combining online reviews with historical sales data. *International Journal of Hospitality Management*, *94*, 102830.

Garnefeld, I., Helm, S., & Grötschel, A.-K. (2020). May we buy your love? psychological effects of incentives on writing likelihood and valence of online product reviews. *Electronic Markets*, 1–16.

Godes, D., Mayzlin, D., Chen, Y., Das, S., Dellarocas, C., Pfeiffer, B., ... Verlegh, P. (2005). The firm's management of social interactions. *Marketing Letters*, *16*(3), 415–428.

Godes, D., & Silva José, C. (2012). The dynamics of online opinion. *Management Science*, *31*(3), 448–473.

Goes, P. B., Lin, M., & Au Yeung, C. M. (2014). "popularity effect" in user-generated content: Evidence from online product reviews. *Information Systems Research*, *25*(2), 222–238.

Homburg, C., Ehm, L., & Artz, M. (2015). Measuring and managing consumer sentiment in an online community environment. *Journal of Marketing Research*, *52*(5), 629–641.

Hu, N., Bose, I., Koh, N. S., & Liu, L. (2012). Manipulation of online reviews: An analysis of ratings, readability, and sentiments. *Decision Support Systems*, *52*(3), 674–684.

Hu, N., Koh, N. S., & Reddy, S. K. (2014). Ratings lead you to the product, reviews help you clinch it? the mediating role of online review sentiments on product sales. *Decision Support Systems*, *57*, 42–53.

Hu, N., Pavlou, P. A., & Zhang, J. J. (2017). On self-selection biases in online product reviews. *MIS Quarterly*, *41*(2), 449–471.

Ireland, R., & Liu, A. (2018). Application of data analytics for product design: Sentiment analysis of online product reviews. *CIRP Journal of Manufacturing Science and Technology*, *23*, 128–144.

Jeong, H.-J., & Koo, D.-M. (2015). Combined effects of valence and attributes of e-wom on consumer judgment for message and product: The moderating effect of brand community type. *Internet Research*, *25*(1), 2–29.

Kim, J., Naylor, G., Sivadas, E., & Sugumaran, V. (2016). The unrealized value of incentivized ewom recommendations. *Marketing Letters*, *27*(3), 411–421.

Kostyra, D. S., Reiner, J., Natter, M., & Klapper, D. (2016). Decomposing the effects of online customer reviews on brand, price, and product attributes. *International Journal of Research in Marketing*, *33*(1), 11–26.

Kumar, N., & Ruan, R. (2006). On manufacturers complementing the traditional retail channel with a direct online channel. *Quantitative Marketing and Economics*, *4*(3), 289–323.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26. doi: doi: 10.18637/jss.v082.i13

Li, X., & Hitt, L. M. (2008). Self-selection and information role of online product reviews. *Information Systems Research*, *19*(4), 456–474.

Li, X., & Hitt, L. M. (2010). Price effects in online product reviews: An analytical model and empirical analysis. *MIS quarterly*, 809–831.

Lin, C. A., & Xu, X. (2017). Effectiveness of online consumer reviews. *Internet Research*, *27*(2), 362–380.

Liu, X., Burns, A. C., & Hou, Y. (2017). An investigation of brand-related user-generated content on twitter. *Journal of Advertising*, *46*(2), 236–247.

Luan, J., Yao, Z., Zhao, F., & Liu, H. (2016). Search product and experience product online reviews: An eye-tracking study on consumers' review search behavior. *Computers in Human Behavior*, *65*, 420–430.

Luke, S. G. (2017). Evaluating significance in linear mixed-effects models in r. *Behavior Research Methods*,

49(4), 1494–1502.

Markinblog. (2021). List of largest ecommerce companies in the world in 2021 (ranked by revenue). *Markinblog*. Retrieved 2021-05-13, from https://www.markinblog.com/largest-ecommerce-companies/

Moe, W. W., & Schweidel, D. A. (2012). Online product opinions: Incidence, evaluation, and evolution. *Marketing Science*, *31*(3), 372–386.

Moe, W. W., & Trusov, M. (2011). The value of social dynamics in online product ratings forums. *Journal of Marketing Research*, *48*(3), 444–456.

Moldovan, S., Goldenberg, J., & Chattopadhyay, A. (2011). The different roles of product originality and usefulness in generating word-of-mouth. *International Journal of Research in Marketing*, *28*(2), 109–119.

Moulton, B. R. (1990). An illustration of a pitfall in estimating the effects of aggregate variables on micro units. *The review of Economics and Statistics*, 334–338.

Mudambi, S. M., & Schuff, D. (2010). Research note: What makes a helpful online review? a study of customer reviews on amazon. com. *MIS quarterly*, 185–200.

Novy-Williams, E., & Soper, S. (2019). Nike pulling its product from amazon in e-commerce pivot. *Bloomberg*. Retrieved 2021-01-20, from https://www.bloomberg.com/news/articles/2019-11-13/nike-will-end-its-pilot-project-selling-products-on-amazon-site

OECD. (2020). E-commerce in the time of covid-19. *OECD*. Retrieved 2021-03-06, from https://read.oecd-ilibrary.org/view/?ref=137_137212-t0fjgnerdbtitle=E-commerce-in-the-time-of-COVID-19

Olagunju, T., Oyebode, O., & Orji, R. (2020). Exploring key issues affecting african mobile ecommerce applications using sentiment and thematic analysis. *IEEE Access*, *8*, 114475–114486.

Park, D.-H., Lee, J., & Han, I. (2007). The effect of on-line consumer reviews on consumer purchasing intention: The moderating role of involvement. *International Journal of Electronic Commerce*, *11*(4), 125–148.

Petersen, M. A. (2009). Estimating standard errors in finance panel data sets: Comparing approaches. *The Review of Financial Studies*, *22*(1), 435–480.

PwC. (2020). Post covid-19 customer strategy implications. *PwC*. Retrieved 2021-03-06, from https://www.pwc.de/de/im-fokus/customercentrictransformation/post-covid-19-customer-strategy-implications.pdf

Rambocas, M., & Pacheco, B. G. (2018). Online sentiment analysis in marketing research: a review. *Journal of Research in Interactive Marketing*, *12*(2), 146-163.

Reisenbichler, M., & Reutterer, T. (2019). Topic modeling in marketing: recent advances and research opportunities. *Journal of Business Economics*, *89*(3), 327–356.

Richter, F. (2021). Nike still on top of the sneaker world. *Statista*. Retrieved 2021-05-13, from https://www.statista.com/chart/13470/athletic-footwear-sales/

Rinker, T. W. (2020). qdap: Quantitative discourse analysis package [Computer software manual]. Buffalo, New York. Retrieved from https://github.com/trinker/qdap (2.4.2)

Rosario, A. B., de Valck, K., & Sotgiu, F. (2020). Conceptualizing the electronic word-of-mouth process: What we know and need to know about ewom creation, exposure, and evaluation. *Journal of the Academy of Marketing Science*, *48*(3), 422–448.

Sahli, R. C. (2020). More people are doing their shopping online and

this trend is here to stay. *CNBC*. Retrieved 2021-03-06, from https://www.cnbc.com/2020/12/15/coronavirus-pandemic-has-pushed-shoppers-to-e-commerce-sites.html

Schuckert, M., Liu, X., & Law, R. (2015). Hospitality and tourism online reviews: Recent trends and future directions. *Journal of Travel & Tourism Marketing*, *32*(5), 608–621.

Schweidel, D. A., & Moe, W. W. (2014). Listening in on social media: A joint model of sentiment and venue format choice. *Journal of marketing research*, *51*(4), 387–402.

Sen, S., & Lerman, D. (2007). Why are you telling me this? an examination into negative consumer reviews on the web. *Journal of interactive marketing*, *21*(4), 76–94.

Sides, R., & Skelly, L. (2021). 2021 retail industry outlook. *Deloitte*. Retrieved 2021-04-13, from https://www2.deloitte.com/content/dam/Deloitte/us/Documents/consumer-business/us-2021-retail-industry-outlook.pdf

Smith, A. N., Fischer, E., & Yongjian, C. (2012). How does brand-related user-generated content differ across youtube, facebook, and twitter? *Journal of Interactive Marketing*, *26*(2), 102–113.

Snijders, T. A. (2005). Fixed and random effects. *Encyclopedia of statistics in behavioral science*, *2*, 664–665.

Snijders, T. A., & Bosker, R. J. (2011). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. Sage.

Statista. (2020). Market share of leading digital camera manufacturers worldwide in 2020, by sales volume. *Statista*. Retrieved 2021-05-13, from https://www.statista.com/statistics/1004962/global-leading-manufacturers-digital-cameras-market-share-sales-volume/

Tsay, A. A., & Agrawal, N. (2004). Channel conflict and coordination in the e-commerce age. *Production and Operations Management*, *13*(1), 93–110.

Wickham, H. (2020). rvest: Easily harvest (scrape) web pages [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=rvest (R package version 0.3.6)

Winter, B. (2013). Linear models and linear mixed effects models in R with linguistic applications. *CoRR*, *abs/1308.5499*. Retrieved from http://arxiv.org/abs/1308.5499

Wu, F., & Huberman, B. A. (2008). How public opinion forms. In *International workshop on internet and network economics* (pp. 334–341).

Xu, G. (2018). The costs of patronage: Evidence from the british empire. *American Economic Review*, *108*(11), 3170–98.

Yang, W., Zhang, J., & Yan, H. (2021). Impacts of online consumer reviews on a dual-channel supply chain. *Omega*, *101*, 102266.

Yang, Y., Mao, Z., & Tang, J. (2018). Understanding guest satisfaction with urban hotel location. *Journal of Travel Research*, *57*(2), 243–259.

Yu, M., Liu, F., Lee, J., & Soutar, G. (2018). The influence of negative publicity on brand equity: attribution, image, attitude and purchase intention. *Journal of Product & Brand Management*, *27*(4), 440–451.

Zhu, F., & Zhang, X. (2010). Impact of online consumer reviews on sales: The moderating role of product and consumer characteristics. *Journal of Marketing*, *74*(2), 133–148.

Zhuang, M., Cui, G., & Peng, L. (2018). Manufactured opinions: The effect of manipulating online product reviews. *Journal of Business Research*, *87*, 24–35.

Zimmerman, B. (2020). Why nike cute ties with amazon and what it means for other retailers. *Forbes*. Retrieved 2021-04-14, from

https://www.forbes.com/sites/forbesbusinesscouncil/2020/01/22/why-nike-cut-ties-with-amazon-and-what-it-means-for-other-retailers/?sh=5ad328c264ff