Building an Online Reputation with Free Content

Entrants into online marketplaces have to build up an online reputation in order to be successful. This problem is especially severe in digital content markets, where consumers are offered an abundance of products that can be characterized as experience goods. I examine how content providers use free content provision as a dynamic pricing strategy to encourage consumers to experience their content, thereby increasing the number of online reviews and contributing towards building up an online reputation. Using data from an online self-publishing e-book platform, I show that although free content provision is effective in increasing the number of online reviews, the same content also receives worse reviews when it is offered for free as compared to being offered at a price. This negative impact of free content provision on online reputation is driven by a selection effect, whereby more consumers with a lower preference for particular content provide reviews when it is offered for free. These results imply that although offering free content can be effective to build up an online reputation more quickly, content providers face the trade-off that the earned reputation will be worse.

Key words: Online Reputation, Online Ratings, Reputation Management, Free Content

1. Introduction

A key institution to reduce information asymmetries and to foster trust in the digital economy are online reputation systems. Since most online marketplaces and platforms rely on these systems, it has become crucial for sellers to establish a track record of positive reviews, that is, they have to establish a good online reputation in order to be successful (Dellarocas, 2003; Chevalier and Mayzlin, 2006). New sellers entering online marketplaces, however, face a start-up problem: As consumers prefer to buy from sellers with an established online reputation, new sellers may find it hard to gather the necessary reviews to establish an online reputation. This problem is aggravated by the public good nature of online reviews resulting from the limited incentives to provide online reviews (Bolton et al., 2004). New sellers therefore face a reputational barrier to entry into online marketplaces – the "cold-start" problem of building a reputation (Milgrom and Tadelis, 2018).

how to success?
=>review-driven
market =>
cold-start =>

previous thoughts

In digital content markets, the cold-start problem of establishing an online reputation is especially severe. Due to 1 the low fixed costs of production and 2 minimal marginal costs of producing and distributing additional content, consumers are offered an abundance of digital content in the long-tail of the market (Anderson, 2004; Brynjolfsson et al., 2006). Additionally, digital content can be characterized as an experience good (Nelson, 1970), which implies that product quality and fit to consumers'

free content provision
to solve cold start

idiosyncratic tastes is uncertain before consumption. This makes it especially hard for new content providers to gain traction in the market, as reviews are underprovided but are especially important to inform consumers about quality and horizontal characteristics of content.

The focus of this article is on free content provision as a strategy for content providers to solve the cold-start problem of building up an online reputation. Providing content for free has the potential to increase the number of consumers who experience the content and who provide online reviews, thereby enabling content providers to build-up a track record of reviews and establish an online reputation more quickly. Therefore free content provision can be understood as a form of dynamic pricing, which has been prominently studied as a strategy by entrants to build a reputation in both the theoretical literature on reputation in Economics (Shapiro, 1983b) and the literature in Marketing studying the relationship between dynamic pricing and word-of-mouth (Kalish, 1983; Ajorlou et al., 2016). In contrast to increasing the number of reviews through fraud (Dellarocas, 2006; Mayzlin et al., 2014; Luca and Zervas, 2016), dynamic pricing is a legitimate way to build up a reputation. It is also more straightforward compared to strategies such as offering financial incentives (Li, 2010; Cabral and Li, 2015; Li et al., 2016) or taking advantage of social interactions and norms to stimulate reviews (Goes et al., 2014; Burtch et al., 2017), which do not solve the problem that consumers need to experience a product in the first place before they can provide credible online reviews. These strategies also often require the platform-owner to implement changes on the platform, whereas using price to stimulate reviews can be directly implemented by sellers and content providers.

compare to other
strategy

My goal is to answer the following questions related to offering free content as a strategy to build up a reputation: (1) Do content providers use free content to build up an online reputation, i.e. are they more likely to offer content for free when it has gathered a lower number of previous reviews? (2) How does offering content for free impact its online reputation, i.e. does free content gather more reviews and are these reviews better or worse compared to the content being offered at a price?

Questions

Prior research does not give a clear answer how offering content for free should affect online reviews and consequently the content's online reputation. Conceptually, the potential effects can occur either *within* or *across consumers*. An *across-consumer* effect occurs when free content is consumed by consumers who provide reviews that are different than those consumers who buy content and provide reviews when it is offered at a price. For example, given downward sloping demand, offering content for free should induce more consumers with a lower preference for the content to acquire it and to provide online reviews. This should result in worse average reviews compared to when the same content is offered at a price. Prior literature has identified similar self-selection effects based on consumer preferences and product experience to be important in the context of online reviews (Li and Hitt, 2008; Dellarocas and Wood, 2008; Moe and Schweidel, 2012; Hu et al., 2017). Therefore, I propose price-induced self-selection as the main mechanism impacting online reputation when content

free-c on consumers:
1. within, 2. cross

is offered for free. However, there are also several mechanisms identified by prior literature that would imply *within-consumer* effects, which occur when the same consumer reviews content differently when she has acquired it for free as compared to when she has purchased the same content at a positive price. For example, if consumers assess product quality ex-post in reference to the purchasing price or in reference to the quality they expected ex-ante based on the purchasing price (Li and Hitt, 2010), 1. the same consumer should provide a better online review when she has acquired content for free. Consumers might also perceive free content to be a gift from the content provider, which they 2 reciprocate by leaving a better online review (Bolton et al., 2004). Research at the intersection of Marketing and Behavioral Economics also suggests that consumers might derive higher actual 3. con benefits from products with a higher price (Shiv et al., 2005), which would predict the same consumer giving a worse review when she has acquired content for free, or that consumers perceive the utility from free content to be higher (Shampanier et al., 2007), which would predict the same consumer giving a better review when she has acquired content for free. In light of these different possible

controversial past evidence

within- and *across-consumer* mechanisms, it is ultimately an empirical question how offering content for free affects its online reputation, which I therefore seek to answer in this paper.

deficiency

Estimating the relationship between offering content for free and its online reputation is challenging. Content quality is usually not observed by the researcher but potentially influences both whether content is offered for free and the number and nature of reviews. This introduces a potential omitted variable bias when trying to establish a causal link between online reputation and offering content for free using observational data. I therefore present an empirical strategy that exploits the fact that I observe in my data the same content at multiple points in time and a subset of this content changes between being offered at a price and being offered for free. As content quality stays fixed over time in my empirical context, I can base my regression analyses only on within-content variation, controlling for content quality using content-level fixed effects.

data source

实例于假设相符合

The empirical setting I study is an online self-publishing platform on which authors regularly offer their e-books as free content. Self-publishing of e-books has become popular due to the low cost of digital content creation and distribution (Waldfogel and Reimers, 2015). Although self-publishing can be an attractive option for authors, as it enables them to either circumvent traditional publishers altogether or to use success in the self-publishing market as a signal to receive more favorable licensing deals (Peukert and Reimers, 2018), the low publishing costs have arguably also resulted in an oversupply of self-published e-books in what can be considered the long-tail (Anderson, 2004; Brynjolfsson et al., 2006) of the e-book market. This makes it very difficult for new self-published e-books by unestablished authors to gain traction in the market. Absent marketing efforts of a traditional publisher, positive online word-of-mouth and reviews are very important for an author to achieve success in the self-publishing market. However, only about 10 percent of e-books on the platform I study have at least one rating, indicating that the cold-start problem of establishing a reputation is especially severe in this market. Offering e-books as free content is a very common strategy by authors to address this problem and is also actively encouraged by the platform that I study.

The results of my empirical analysis show that consistent with free content provision being used as a strategy by unestablished authors to build up a reputation, an increase in the number of reviews by 100 percent decreases the probability that an e-book is offered as free content by 15 to 17 percent. Depending on the sub-sample of e-books studied, an e-book also receives between 1 percent and 5 percent more reviews when it is offered as free content. However, offering an e-book as free content lowers its ratings by 0.05 to 0.06 stars (out of 5 possible stars), corresponding to a decrease of 6 to 7 percent of the standard deviation of average ratings. At the same time, offering an e-book as free content increases the standard deviation of ratings by 5 to 6 percent and results in review text for free e-books that is shorter and written in a worse sentiment. I show that these effects are explained by the hypothesized *across-consumer* mechanism based on price-induced self-selection of consumers, as I show that reviewers of e-books that are offered as free content are substantially different in terms of their observable characteristics compared to reviewers of e-books that offered as paid content, which suggests they have a lower preference for the reviewed e-books.

con 1

con 2

These results have the following implications for content providers and content platforms: For providers, they imply that although offering free content can be an effective strategy to increase the number of reviews and to build up an online reputation more quickly, it involves a trade-off, as it also results in worse reviews and increases the standard deviation of ratings. As this is driven by a selection effect whereby free content attracts consumers with lower preferences, it suggests that content providers could actively manage this trade-off, for example by targeting free content at consumers who are more likely to have a higher preference for the content. For content platforms, the results imply that the entry barrier of needing to build up an online reputation cannot be solved by content providers using pricing alone. Ignoring this problem could lead to lower entry of content providers, decreasing the variety of content on a platform, thereby reducing the attractiveness of the content platform for consumers and thus endangering its success.

pros and cons

explain

solve to cons

2. Literature

This article contributes to several strands of literature in Information Systems, Economics and Marketing. In particular, it contributes to the literature on online reviews and reputation, which studies the impact of online reviews, biases in online reviews and how firms can manage their online reputations, but also to the literatures on dynamic pricing, free digital content and free sampling.

2.1. Online Reputation Management

In recent years, a growing literature in Information Systems (e.g. Ba and Pavlou, 2002; Li and Hitt, 2008; Forman et al., 2008), Economics (e.g. Lucking-Reiley et al., 2007; Resnick and Zeckhauser, 2002; Cabral and Hortacsu, 2010) and Marketing (e.g. Chevalier and Mayzlin, 2006) has studied online reputation systems. Surveys can be found in Dellarocas (2003), Cabral (2012a), Tadelis (2016) and Luca (2017). In general, the literature has found a positive impact of sellers' online reputations on their success in online marketplaces.

reputation => better performance.

Based on these findings, a follow-up literature has examined how sellers can manage their online reputation. A particular concern is that online sellers might use strategies that undermine the effectiveness of online reputation systems, such as sellers leaving positive reviews for their own or negative reviews for their competitors' products or paying third parties to do so (Dellarocas, 2006; Mayzlin et al., 2014; Luca and Zervas, 2016), or changing their online identities (Friedman et al., 2007; Wibral, 2015). However, there is also a broad literature studying legitimate ways how sellers can manage their online reputation (e.g. Goes et al., 2014; Burtch et al., 2017; Proserpio and Zervas, 2017). In particular, Li (2010), Li and Xiao (2014), Cabral and Li (2015), Fradkin et al. (2016), and Li et al. (2016) show that platforms and sellers can incentivize buyers to provide reviews by offering them rebates for leaving reviews. In terms of methodology, this article is particularly close to Li et al. (2016), who use observational data from an online marketplace to study sellers' decisions to adopt a rebate mechanism to induce more reviews. However, an important difference between offering monetary incentives in the form of rebates and offering a product for free is that a rebate for providing a review is typically offered after a consumer purchases a product.[1] Therefore it is not clear whether the results from these studies would carry over to the case of offering free content as a way of stimulating reviews.

manage it!

other's method and result might not applicable here.

2.2. Biases in Online Reviews

My paper is also related to the literature studying potential biases of online reviews. In particular, Chevalier and Mayzlin (2006, p.345) already raise the point that there is a potential selection bias in online reputation systems since only consumers who expect to enjoy a product will choose to acquire it, which makes the pool of reviewers different from the general population.

The literature has established different forms of biases in online reviews, such as reciprocity bias (Bolton et al., 2004, 2013; Fradkin et al., 2016), selection bias (Li and Hitt, 2008; Gao et al., 2015; Hu et al., 2017; Brandes et al., 2018), biases induced by price effects (Li and Hitt, 2010), or biases induced by herding or social cues (Lee et al., 2015; Godinho de Matos et al., 2016; Wang et al., 2018). I contribute to this literature by showing that a zero price can induce a selection effect that

[1] A rebate would only induce the same effect on reviews as offering a product for free if all consumers would receive the rebate and would be ex-ante informed about the rebate.

zero price decrease online rateing(biase!)

how I solve biases

biases online ratings downwards. A conceptually similar bias is the "Groupon effect" found by the Information Systems literature studying discount offers on platforms such as *Groupon.com* (Byers et al., 2012; Li, 2016; Chaudhari and Byers, 2017), which has found that discount can have a negative effect on online ratings. This can also be explained by a selection of different consumers acquiring a product when it is promoted with a discount, but it is not clear whether selection is induced only by the price-discount or by the accompanying sales promotion run on the discount-platforms.

discount effect

mutual rich

A particularly relevant study to this work is Lin et al. (2019), who study how the provision of free samples impacts online reviews in the context of physical products offered on an e-commerce platform. They find that free samples generate better reviews of the sampled products due to a reciprocity effect. In the context of free content provision, this article does not find evidence for a positive reciprocity effect but finds that free content provision generates worse reviews due to a selection effect. Therefore, this work can be seen as extending the work of Lin et al. (2019) into the context of digital content, for which consumer tastes are more heterogeneous and uncertainty about product characteristics higher, which increases the likelihood of a negative selection effect occurring. In particular, there are no direct production costs for content providers to offer free content, which can explain why free content provision does not trigger reciprocity as in the case of free samples of physical products that are costly to provide for the seller.

2.3. Dynamic Pricing and Reputation

This article examines free content provision as a strategy for building up a reputation in the early stages of the life-cycle of content. This can be understood as a form of dynamic pricing whereby the content provider offers content initially for free and after earning a reputation switches to offering content at a price.[2] Therefore, this work is related to the literature studying the interplay between dynamic pricing and word-of-mouth (Kalish, 1983, 1985). In particular, Juan et al. (2019) examine how sellers' pricing strategy is impacted by online reviews. My results add to this literature by showing that a more extreme dynamic pricing strategy in the form of offering a product for free can be disadvantageous for products where consumer tastes are heterogeneous such as for digital content.

There is also an extant theoretical literature in Economics studying reputation. Literature reviews are provided by Cabral (2005); Mailath and Samuelson (2006); MacLeod (2007); Cripps (2009); Bar-Isaac and Tadelis (2008). In particular, Bergemann and Välimäki (2000, 2006) study the incentives of competing sellers to lower their prices to encourage experimentation and learning of their products' quality. Other similar models are McFadden and Train (1996) and Bose et al. (2006). However, these articles typically assume homogeneous preferences of buyers regarding the type of the seller which rules out any selection effects. Two recent exceptions are Bar-Isaac and Deb (2014b,a), who, however,

[2] Nair (2019) reviews the literature in Marketing and Economics studying pricing over a product's life cycle.

do not consider pricing in their models. Therefore, this study highlights a trade-off that to-date is mostly absent from the theoretical literature on reputation in Economics.[3]

2.4. Free Digital Content and Sampling

Finally, this article contributes to the literatures on free digital content and sampling. Much of the literature on free digital content focuses on how digital content providers can monetize their content, for example by converting unpaid users to paid users by using the "freemium" business model (Anderson, 2009; Pauwels and Weiss, 2008; Oestreicher-Singer and Zalmanson, 2013; Bapna and Umyarov, 2015; Bapna et al., 2018; Runge et al., 2019) or by using free content to attract eyeballs for advertisers (Lambrecht and Misra, 2017). In particular, I contribute to the literature examining how free content can be used as a promotional tool (Dewan and Ramaprasad, 2012, 2014; Goh et al., 2013; Zhang, 2018) showing how offering free content can be used to establish a reputation that can be monetized in later stages of the life-cycle of content.

monetize

There are some studies looking at online reviews and online word-of-mouth in the context of free content provision. In particular, Claussen et al. (2013) show how user engagement is impacted when product ratings become more salient in the context of free Facebook. Oh et al. (2016) examine the effect of a paywall versus free content provision on online word-of-mouth in the context of online news content. Choi et al. (2019) examine how free samples and reviews interact to influence a consumer's purchasing decision in the context of digital content in the form of e-books.

other content

There is also an extant literature focusing on substitution between free and paid content in the form of pirated content (Chellappa and Shivendu, 2005; Smith and Telang, 2009; Johar et al., 2012; Aguiar et al., 2018) or legally available content (Xu et al., 2014; Kretschmer and Peukert, forthcoming), and literature on free sampling of both physical products in offline contexts (Marks and Kamins, 1988; Heiman et al., 2001; Bawa and Shoemaker, 2004) as well as on digital products in online contexts (Wang and Zhang, 2009; Cheng and Liu, 2012; Lee and Tan, 2013).

3. Theory and Hypotheses

In this section, I present a theoretical framework to derive hypotheses that will guide my empirical analyses. The framework distinguishes between two cases: offering content for free or at a positive price. The framework addresses two related questions: When will a content provider offer her content for free with the goal of building up a reputation and how will offering the content for free impact the online reviews it receives?[4]

Q

[3] Two recent exceptions are working papers by Acemoglu et al. (2017) and Stenzel et al. (2019).

[4] I distinguish only between a positive price and a zero price to be consistent with the empirical setting of free content provision. However, all results carry over to a framework where a seller can set a high or low but non-zero price.

Consider a market where a content provider offers her content at multiple time periods $t \in \{1, 2, 3, ..., T\}$. The content has marginal production costs of zero and in each period the content provider can choose to offer the content either as free content with a price $p_t = 0$ or as paid content at some positive price $p_t > 0$. In each time period t, a new consumer i is deciding whether to acquire the content with his utility function for consuming the content given by

$$u_i(q, \epsilon_i), \tag{1}$$

which is assumed to be increasing in the content's quality q and the consumer's idiosyncratic taste for the content ϵ_i. The taste parameter ϵ_i is assumed to be identically and independently randomly iid distributed in the population of consumers. Quality q of the content is the same for all consumers but initially unknown to them. They only know the distribution of feasible qualities and can infer quality from observing previous reviews summarized in the online reputation of the content.

Before acquiring the content, a consumer is not certain about the utility he will receive from consuming the content but receives a noisy signal s_i of his utility and observes the content's online 我认为这个内容如何 reputation $R_t = \{r_1, ..., r_{t-1}\}$ consisting of reviews $r_1, ..., r_{t-1}$ written by consumers who acquired and consumed the content in previous time periods. The noisy signal s_i is assumed to be increasing in the consumer's utility such that $E(s_i|u_i') > E(s_i|u_i'')$ for any two utilities $u_i' > u_i''$. The content therefore has both characteristics of a search and an experience good (Nelson, 1970; Li and Hitt, 2008).[5] After observing the signal s_i and the content's online reputation R_t, the consumer forms a conditional expectation $E(u_i|s_i, R_t)$ of his utility for the content and acquires it if[6]

$$E(u_i \mid s_i, R_t) > p_t. \tag{2}$$

price

If the consumer chooses to acquire the content, he learns his actual utility after consuming the content and with some constant probability leaves a review. The review consists of a rating that is equal to the consumer's utility, i.e.

$$r_t = u_i(q, \epsilon_i). \tag{3}$$

and a written review that reflects the consumer's utility from consuming the product. This implies that for an outside observer, a better review implies a higher likelihood that the content is of high

[5] In the model setup of Li and Hitt (2008), a good has both "search attributes," which can be inspected by consumers before purchasing, and "experience attributes," which consumers learn only after consuming the product. See also Villas-Boas (2004) and Hu et al. (2017) for models with a similar set-up.

[6] As the consumer's expectation does not include considerations of the content provider's equilibrium strategy in current and previous periods and how it is related to the content's quality, consumers can be considered to be boundedly rational in the context of this model. Furthermore, I assume that consumers do not take into account potential biases in previous online reviews R_t driven by the content provider's past prices. Hu et al. (2017) provide empirical evidence in favor of this assumption in the context of self-selection biases in online review systems.

quality but the review is conflated with consumer i's idiosyncratic taste ϵ_i. Notice also that the rating r_t does not directly depend on the content's price p_t, which implies that the consumer assesses the content only based on how much he enjoyed consuming it and does not take into consideration how much he paid or how much he expected to enjoy it.[7] As the probability of observing a review from a given consumer is conditioned on the consumer acquiring the content, and a lower price increases the probability that the condition in equation (2) is fulfilled, the probability to observe a review increases when content is offered for free, which implies the following hypothesis:

absurd ?

Hypothesis 1. Content will receive more reviews when it is offered for free ($p_t = 0$) as compared to when it is sold at a price ($p_t > 0$).

E(U | R, s) >= P

As a better online reputation R_t increases the conditional expectation of consumers' utility in equation 2 and thus the probability that they acquire it, in any time period $t < T$ a content provider has to consider whether to offer the content at price $p_t = 0$ to increase the likelihood to receive an additional online review. Choosing $p_t = 0$ implies sacrificing potential current profits to increase the likelihood of receiving an additional online review and potentially earning higher profits in all future time periods due to a better reputation. Such an investment in future reputation $R_{t+1}, ..., R_T$ is generally more profitable when content has received a lower number of previous reviews as then an additional review has a larger impact on a consumer's expectation $E(u_i \mid s_i, R_t)$. This is a common result in the theoretical literature in Economics on reputation building (Shapiro, 1983b,a; Bar-Isaac and Tadelis, 2008) and has been applied in the context of online reputation and conditional rebates (Li et al., 2016). In a similar vein, the theoretical literature in Economics on advertising finds that a firm should invest more in brand awareness in the early stages of a product's life cycle (Milgrom and Roberts, 1986; Bagwell, 2007). Although this literature also suggests that the incentives to invest in reputation are higher for high-quality sellers (e.g. Nelson, 1974; Kihlstrom and Riordan, 1984; Milgrom and Roberts, 1986), the literature on observational learning (Banerjee, 1992; Bikhchandani et al., 1992; Anderson and Holt, 1997; Celen and Kariv, 2004; Liu, 2006; Dellarocas et al., 2007; Duan et al., 2008) has found that the mere number of observable previous purchases can already increase the demand for a product, as consumers infer a higher quality from observing a larger number of previous consumers purchasing a product or that more previous purchases indicate that a product has a broader general appeal in terms of its horizontal characteristics (Tucker and Zhang, 2011). Taken together, this implies that providers of low-quality content also have an incentive to invest in gathering more online reviews, which is also consistent with prior literature finding that additional

[7] An alternative to this assumption would be that consumers directly rate content in reference to its sales price, i.e. $r_i = u_i(q, \epsilon_i) - p_t$ or that they rate the content comparing it to their expected utility for which they might use the sales price to infer the content's quality, i.e. $r_i = u_i(q, \epsilon_i) - E(u_i|s_i, R_t, p_t)$. Li and Hitt (2010) in particular explore a similar model set-up.

好的坏的都要打广告， 因为评价数量才是第一生产力！

to the average rating, the total number of previous reviews has a positive impact on sales (Liu, 2006; Dellarocas et al., 2007; Duan et al., 2008). This is summarized in the following hypothesis:

Hypothesis 2. Content will be more likely to be offered for free ($p_t = 0$)*, the lower the number of previous reviews of the content.*

E(U | R, s) >= P

As the probability of observing a review depends on a consumer's expectation of his utility to exceed the content's price, and both the rating a consumer assigns in her review and her expectation are increasing in her actual utility, the expected value of an observed rating should be increasing in the price of the content, or more formally:

$$\frac{\partial E(r_i|p_t)}{\partial p_t} > 0. \tag{4}$$

selection effect fomula, higher price => better rating

The intuition is that a consumer with a higher personal taste ϵ_i will in expectation receive a higher signal s_i and thus be more likely to acquire the content and to provide an online review. This results in a selection effect in the form that the higher the price p_t, the higher the expected value of the taste parameter ϵ_i of the consumer who acquire the content and who consequently provides review. This implies the following hypothesis:

Hypothesis 3. Content will receive on average worse reviews when it is offered for free ($p_t = 0$) *as compared to when it is offered at a positive price* ($p_t > 0$).

An additional testable implication of the selection effect is that a lower price, implying a weaker selection of consumers based on their expected utility, will increase the dispersion of observed ratings. The reason is that a lower price makes it more likely that both consumers with a high and a low personal taste ϵ_i receive signals of their utility such that their expected utility exceeds the content's price, which implies

$$\frac{\partial Var(r_i|p_t)}{\partial p_t} < 0 \tag{5}$$

variation increased when lower price

and the following hypothesis:

Hypotheses 4: The variance of the ratings of content will be larger when it is offered for free ($p_t = 0$) *as compared to when it is offered at a price* ($p_t > 0$).

4. Empirical Context and Data

4.1. About Online Self-Publishing and Smashwords

I use data collected from the online self-publishing platform *Smashwords.com* to test the hypotheses derived in the previous section. While initially a niche-market, direct self-publishing of e-books on platforms such as *Smashwords* has evolved in recent years to a market of substantial size. The growth of self-publishing has been mainly triggered by a decrease in production and distribution costs to virtually zero due to advances in digital publishing and the introduction and diffusion of digital e-readers such as *Amazon's Kindle*. By 2013, the market share of self-published e-books had already grown to an estimated tenth of both the number of books in bestseller lists and overall unit sales (Waldfogel and Reimers, 2015).

Smashwords is one of the largest online distributors of self-published e-books (Bowker, 2014). E-books that are published on *Smashwords* are sold on its website and also distributed to the largest online e-book retailers such as *Apple's iBookstore, Barnes & Noble,* or *Kobo.* Most titles are published by authors who can be considered amateurs, similar to content providers on other digital platforms (Boudreau, 2018). Although most titles on *Smashwords* only serve small niche segments of the market, some titles have also been very successful, even reaching international bestseller status.[8] Data collected from *Smashwords* has been used in previous research, such as in Peukert and Reimers (2018), studying the effects of digital disintermediation on the supply of new products and Zegners (2016), studying the relationship between pricing and information disclosure in the form of digital samples.

why this data

The following features make the online self-publishing market a good setting to examine the strategy of giving away free content with the goal of establishing a reputation: Whereas authors who release their books via traditional publishers are supported by publishers, who run advertising campaigns, induce newspapers to publish reviews, or organize book tours, self-published authors cannot make use of these complementary services provided by publishers. To inform potential readers of their books' existence, characteristics, and quality; self-published authors have to rely on social media, word-of-mouth, and online reviews and ratings. Offering e-books as free content is a particularly popular strategy to reach these goals. Often, authors either temporarily offer e-books for free by setting their price to zero, or they offer certain e-books permanently for free, for example, the first volume of a series of e-books. *Smashwords* actively encourages this strategy, arguing that it helps to gather first reviews and to build up a readership as an author.[9]

4.2. Datasets

I use two complementary datasets in my empirical analyses. First, I have collected via a web-scraper data on the whole population of e-books available on *Smashwords* at four points in time: June 2015, February 2016, January 2017 and July 2018.[10] The data contain each e-book's title, author, date of

[8] Coker, M. (2012). "How to Sell E-books at the Apple iBookstore." *Smashwords Blog, Available at http://blog.smashwords.com/2012/11/how-to-sell-ebooks-at-apple-ibookstore.html, accessed on January 16, 2019.*

[9] See for example: http://blog.Smashwords.com/2015/08/how-to-price-kindle-books-to-free.html, accessed on February 15, 2019).

[10] See figure 2 in the Online Appendix for a screenshot of an e-book's website on *Smashwords* from which the data were collected.

publishing, price, genre and subgenre, number of reviews and average rating. I label e-books with a price of zero as free e-books. As many authors offer multiple e-books, I also compute the number of e-books published by an author, the overall number of reviews for all e-books of an author, and the number of e-books an author offers as free content. All of these variables are recomputed for every wave in my data. I also exclude 65,000 e-books from my dataset of authors who have published more than 50 e-books, as they represent only 0.4 percent of authors publishing on *Smashwords* but supply 16 percent of all e-books. Excluding these authors has only an impact on the summary graphs and statistics but not on the main results of the empirical analysis.[11]

数据清洗

The final dataset has a panel structure, with 1,086,266 observations on 361,206 unique e-books. This data will be used to examine whether offering an e-book as free content increases the number of reviews it receives (Hypothesis 1) and whether an e-book with a lower number of reviews is more likely to be offered as free content (Hypothesis 2). Table 1 shows the summary statistics of the panel dataset. About 18 percent of observations in the panel are on free e-books. Although a small number of e-books have extremely high prices, such as one particular e-book that is offered at a price of $100 Mio., the median price of an e-book is $2.99.

The second dataset is taken from the last wave of data collected and contains information on each rating and review that has been given to an e-book up to July 2018. As can be seen in figure 3 in the Online Appendix, each e-book's website on *Smashwords* shows ratings and reviews of the e-book, including the text of the review, the name of the reviewer, the date of the review and whether the review was given for a free or a paid version of the e-book. Additionally, I compute for each review the number of previous reviews for the same e-book and the number of characters contained in the review text. The information on whether a rating is given for a free e-book will be crucial for identifying the effect of offering an e-book for free on a rating and the accompanying written review (Hypothesis 3) and the dispersion of ratings (Hypothesis 4). Overall, 48 percent of reviews in the dataset were given to free e-books.

4.3. Identification

The main identification strategy in this article is to exploit the fact that in both datasets the same e-books and their reviews are observed at multiple points in time. Thereby e-book level fixed effects can be included in all regression analyses, which allows the hypothesized effects to be estimated only through within e-book variation. This exploits the fact that there is a subset of e-books whose authors change their prices from zero to a positive price or vice-versa. Such an approach that uses only within product or within seller variation has been used in previous studies on online reputation,

[11] The most extreme case is one author publishing 17,000 musical scores of well-known tunes and classical pieces.

Table 1 Summary Statistics at the E-book Level

Statistic	N	Mean	St. Dev.	Min	Median	Max
Free e-book (1=yes)	1,086,266	0.18	0.38	0	0	1
Price	1,086,266	310.40	161,832.00	0.00	2.99	100,000,000.00
Number Reviews	1,086,266	0.32	1.99	0	0	428
Average rating	144,445	4.45	0.81	1.00	5.00	5.00
Weeks since published	1,086,266	158.10	99.41	0.14	149.40	539.90
Number e-books same author	1,086,266	9.49	10.78	1	5	50
Number reviews other e-books by author	1,086,266	2.54	12.32	0	0	1,097
Number free e-books by same author	1,086,266	1.49	4.10	0	0	49

Note: This table shows summary statistics on the level of an e-book. All four waves of scraped data from *Smashwords* are included, therefore the same e-book can be included up to 4 times in the dataset.

Table 2 Summary Statistics at the Rating Level

Statistic	N	Mean	St. Dev.	Min	Median	Max
Rating (1-5 stars)	99,379	4.51	0.85	1	5	5
Free e-book reviewed (1=yes)	99,379	0.48	0.50	0	0	1
Number weeks reviewed after published	99,379	36.14	59.01	−0.14	10.29	530.10
Number previous reviews	99,379	5.96	24.00	0	1	427
Number characters in review	99,379	473.50	705.50	1	240	27,572

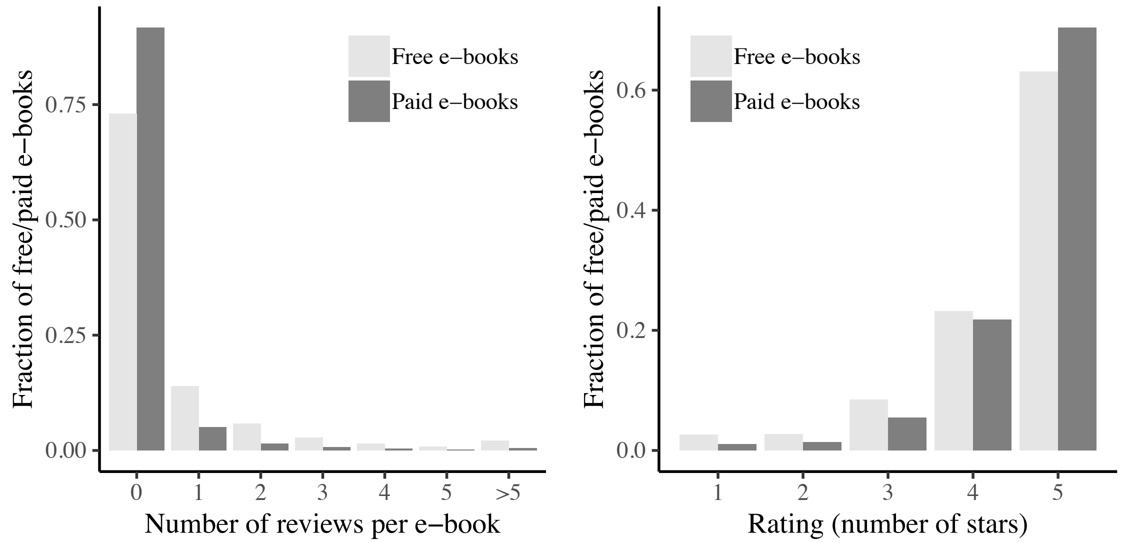
Note: This table shows summary statistics on the level of each rating. The data on ratings are taken from the last wave scraped from *Smashwords* in July 2018.

e.g. in Chevalier and Mayzlin (2006), Cabral and Hortacsu (2010), and Elfenbein et al. (2012, 2015). In particular, the approach to identifying changes in content providers' strategies exploiting the panel structure of the data is inspired by Li et al. (2016), who use data on sellers from an online platform who adopt "Rebate-for-Feedback" mechanisms. They identify the effects of rebates on the nature of feedback using within-seller changes in rebate-adoption. In a similar vein, Einav et al. (2015, 2018) use matched listings of the same seller to assess the causal effect of sales strategies in online markets.

control on unobservable error

The main remaining endogeneity concern is that unobserved time-varying factors influence both the decision of an author to offer an e-book as free content and its online reviews. While I cannot fully rule out such concerns, all of my regressions include the time an e-book has spent on the market as a control for general trends on the level of an e-book. For example, consumers with differing preferences might acquire an e-book at different points in time (Li and Hitt, 2008). Additionally, I include month-year fixed effects and wave (in which data where collected) fixed effects in all specifications to control for seasonal effects and overall changes on the platform level, such as the number of e-books offered and different cohorts of authors and consumers entering the platform over time.

Figure 1 Distribution of Ratings and Number of Reviews for Free and Paid E-Books



Note: The left graph uses data on the e-book level from the last wave of data collection in July 2018. The right graph shows the distribution of ratings including all 99,379 observations in the dataset.

5. Results

Feb 10, 11:13 Paris

In the following results section, I first present analyses on the e-book level, testing whether an e-book receives more reviews when it is offered as free content (Hypothesis 1) and whether an e-book is more likely to be offered as free content when it has gathered a lower number of previous reviews (Hypothesis 2). In a second step, I present analyses on the level of individual ratings, testing whether an e-book receives worse ratings (Hypothesis 3) and ratings with a higher standard deviation (Hypothesis 4) when it is offered as free content. Finally, I show results providing evidence that the observed effects on reviews are driven by a selection effect induced by e-books' price as hypothesized in the theoretical section.

5.1. Analysis on E-book Level

5.1.1. Free Content Receive More Reviews. According to Hypothesis 1, content should receive more reviews when it is offered for free as compared to when it is offered at a price. The left panel of figure 1 shows the raw distribution of the number of reviews per e-book, splitting the data into e-books offered as free content and e-books offered at a price. Consistent with Hypothesis 1, e-books that were offered as free content had gathered overall more reviews than e-books offered at a price. While only 8 percent of priced e-books had received at least one rating, 27 percent of free e-books had received at least one rating.

To examine the relationship between whether an e-book is offered as free content and the number of reviews it receives in more detail, in particular controlling for observed time-varying and unobserved

Table 3 Impact of Offering E-Book as Free Content on Number of Reviews

	Dependent variable:					
	Log(N Reviews, t)					
	(1)	(2)	(3)	(4)	(5)	(6)
Free e-book (Yes=1), t-1	0.011***	0.011***	0.011***	0.008***	0.006*	0.047**
	(0.002)	(0.002)	(0.002)	(0.002)	(0.003)	(0.020)
Weeks since published, t		0.0001***	0.00005***	0.00004***	0.0001***	0.003***
		(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00005)
Log(N e-books same author, t-1)			0.009***	0.006***	0.013***	0.016
			(0.002)	(0.001)	(0.004)	(0.026)
Log(N reviews other e-books same author, t-1)			0.060***	0.020***	0.020***	−0.015
			(0.004)	(0.003)	(0.005)	(0.013)
Log(N other free e-books same author, t-1)			0.0001	0.002*	0.001	0.072***
			(0.001)	(0.001)	(0.003)	(0.022)
Log(Number reviews, t - 1)				0.312***	0.360***	
				(0.009)	(0.015)	
Average rating, t-1					−0.006	
					(0.009)	
Fixed effects e-book	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
Fixed effects time-period	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
Observations	725,060	725,060	725,060	725,060	101,625	17,173
Number e-books	290,681	290,681	290,681	290,681	38,002	6,726

*p<0.1; **p<0.05; ***p<0.01
Clustered standard errors in parentheses

Notes: Standard errors are clustered on the e-book level. Column (5) includes only e-books that have at least one rating and where the average previous rating can be computed. Column (6) includes only e-books where the number of reviews increases during the time-frame the data were collected. To all variables where the logarithm is taken one is added to account for observations that are zero.

time-constant differences between free and priced e-books, I continue by estimating the following regression equation:

$$Log(NReviews_{it}+1) = \beta_1 FreeBook_{i,t-1} + x'_{i,t-1}\boldsymbol{\beta} + b_i + \tau_t + \epsilon_{it}, \tag{6}$$

where $NReviews_{it}$ is the total number of reviews e-book i had received by time t, $FreeBook_{i,t-1}$ is a binary indicator of whether an e-book is offered for free at time $t-1$, $x'_{i,t-1}$ are time-varying controls, b_i and τ_t are e-book and time-period fixed effects, and ϵ_{it} is the error term. Time-periods t denote in which of the four waves an observation was collected. As the e-book level fixed effects b_i pick up any time-invariant differences between e-books such as their quality, author, genre etc., the coefficients in this model are identified only through within e-book variation, i.e. through e-books where changes in the independent variables are observed. As the distribution of the variable is heavily skewed, I use the logarithm of the number of reviews per e-book, adding 1 to the number of reviews to account for e-books with zero ratings. I use the same transformation throughout the paper for all other variables that are similarly skewed, such as the total number of e-books and reviews per author. I use lagged values of the independent variables as I do not observe at which exact time the value of a variable changes within the time period that passes between the collection of two subsequent waves.[12] Standard errors ϵ_{it} are clustered on the e-book level.

Table 3 shows the results of a series of regression models, subsequently adding a number of controls. The coefficient on whether an e-book is offered for free is positive and statistically significant in all specifications, thus confirming Hypothesis 1. The size of the coefficient does not substantially change when controlling for the number of weeks an e-book has been offered on the platform in column (2); the overall number of e-books the same author offered on the platform, how many of these e-books were also offered for free and the total number of reviews these e-books had received in column (3); the number of reviews of the same e-book in the previous time period in column (4) and the average of previous ratings in column (5). The size of the coefficient implies that an e-book that is offered as free content receives 0.6 to 1.1 percent more reviews per time period as compared to the case where the same e-book is offered at a price.

Although the results confirm Hypothesis 1, it is notable that the size of the effect is small. One possible explanation is that since there is a very large number of e-books being offered as free content on the platform, over 68,000 e-books where offered as free content at the time the last wave of data was collected, any single e-book has only a small probability to be discovered and to be reviewed by readers, even if it is offered for free. While I cannot test this explanation explicitly, an indirect test is to exclude e-books from the regression for which no change in the number of reviews was observed within the time-frame the data were collected. These e-books are likely to have escaped the attention of readers. Excluding these e-books from the regression in column (5) indeed increases the size of the effect of whether an e-book is offered for free substantially, indicating that in this subset of the data an e-book receives 4.7 percent more reviews ($p < 0.05$) when it is offered as free content.[13] Although this result needs to be interpreted with caution, as excluding observations where the dependent variable does not vary within groups naturally increases the size of the estimated coefficients, one plausible conclusion is that an important boundary condition for free content provision to increase the number of reviews is that the content offered for free needs to gain the attention of consumers in the first place.

Looking at the impact of the control variables, the positive coefficients on the number of other e-books and the overall number of reviews for these e-books in columns (3) and (4) indicate the presence of spill-over effects between e-books, as an e-book receives more reviews when its author has published overall more e-books on the platform and when these e-books have received more reviews. The positive ($p < 0.01$) and substantive size of the coefficient on the number of reviews for

[12] The only variable that is included with non-lagged values is the variable that measures how many weeks have passed since an e-book has been published. This variable mechanically increases during the time that passes between the collection of two waves and therefore there cannot be any reverse causality driving changes in the variable.

[13] Commonly used approaches to estimate non-linear regression models with group-fixed-effects such as Conditional Logit similarly drop groups with no variation of the dependent variable from the estimation (Beck, 2018).

the same e-book in the previous period in column (4) also highlights that it is crucial for authors to establish an online reputation in the first place, as once authors have gathered their first reviews for an e-book it becomes easier to build on this online reputation and gather more reviews. The average previous rating on the other hand does not have a statistically significant impact on the number of reviews, which implies it is important to establish a reputation but not necessarily to establish a "good" reputation.

As a robustness check, table 12 in the Online Appendix shows regression results where only e-books that exhibit variation in the variable indicating whether an e-book was offered as free content are included. This reduces the sample size to $25,243$ observations on $8,763$ e-books. The results and the significance level are qualitatively similar.[14] Table 13 in the Online Appendix shows regression results using a balanced panel of $206,567$ e-books that were observed in each of the four time periods the data was collected. The coefficients and their significant levels remain virtually unchanged.

5.1.2. Content with Fewer Reviews is More Likely to be Offered for Free. In this section, I test Hypothesis 2 according to which content should be more likely to be offered for free when it has gathered a lower number of reviews because content providers then face stronger incentives to invest into their reputation by offering their content for free. I continue by estimating the following regression equation:

$$\begin{aligned} FreeBook_{it} =& \beta_1 NReviews_{i,t-1} \\ & + x'_{i,t-1}\boldsymbol{\beta} + b_i + \tau_t + \epsilon_{it}, \end{aligned} \tag{7}$$

where $FreeBook_{it}$ is a binary indicator of whether e-book i was offered as free content at time period t, $NReviews_{i,t-1}$ is the number of reviews of e-book i in time period $t-1$, $x'_{i,t-1}$ are time-varying controls, b_i and τ_t are e-book and time period fixed effects, and ϵ_{it} is the error term. As in the previous analysis, the e-book level fixed effects b_i pick-up any time-invariant differences between e-books. The coefficient β_1 is only identified through within e-book variation in the number of reviews and therefore measures whether an increase in the number of reviews of a particular e-book is followed by a change in whether the same e-book is offered as free content. The results shown in table 4 are estimated using a Conditional Fixed Effects Logit model and coefficients are computed as average marginal effects.[15] As computing marginal effects in the case of a Fixed Effects Conditional Logit Model requires assuming that the group-level fixed effects are equal to zero, the reported coefficients should be interpreted as changes in the probability to be offered for free in the case of an e-book that is offered at a positive price throughout the observation period.

[14] As e-books that exhibit no variation in being offered for free still contribute towards estimating the coefficients on the control variables, the main specifications in table 3 that include all e-books are my preferred models.

[15] Table 14 in the Online Appendix shows results estimated with a Linear Probability model. Most of the coefficients have the same sign and the main effect on the number of reviews is also negative ($p < 0.05$). The effect size is, however,

Table 4 Drivers of Offering E-book as Free Content - Logit Model

	Dependent variable:			
	Free Book (Yes=1), t			
	(1)	(2)	(3)	(4)
Log(Number reviews, t-1)	−0.172***	−0.172***	−0.151***	−0.179
	(0.0353)	(0.0352)	(0.0344)	(0.144)
Weeks since published, t		−0.000187***	−0.000166***	−0.0000278
		(0.0000439)	(0.0000437)	(0.0000843)
Log(N e-books same author, t-1)			0.0627**	−0.00980
			(0.0191)	(0.0040)
Log(N reviews other e-books same author, t-1)			−0.0405	−0.0206
			(0.0222)	(0.0412
Log(N other free e-books same author, t-1)			0.00207	0.0693
			(0.00578)	(0.003)
Average rating, t-1				0.0741*
				(0.0329)
Fixed effects e-book	*Yes*	*Yes*	*Yes*	*Yes*
Fixed effects time-period	*Yes*	*Yes*	*Yes*	*Yes*
Number e-books	9,282	9,282	9,282	2,113
Observations	26,905	26,905	26,905	6,181

*p<0.1; **p<0.05; ***p<0.01
Standard errors in parentheses

Note: This table shows results from Conditional Logit regressions with fixed effects. Shown coefficients are computed as average marginal effects. Column (4) only includes e-books for which at least one rating is available. To all variables where the logarithm is taken one is added to account for observations that are zero.

The results in table 4 generally confirm Hypothesis 2: In columns (1) to (3), the number of previous reviews has a negative impact ($p < 0.01$) on whether an e-book is offered as free content. Only in column (4), where additionally the average of previous ratings of an e-book is included as a control, the coefficient on the number of previous ratings is not statistically significant. This is possibly explained by the loss in statistical power due to the exclusion of all observations where an e-book had not gathered any previous reviews. The size of the coefficients indicates that the probability of offering an e-book as free content decreases by 15 to 17 percent after an increase in the number of previous reviews by 100 percent.

Looking at the control variables, the number of other e-books that an author has published on the platform has a positive impact on whether an e-book is offered as free content in column (3) ($p < 0.05$), which is consistent with authors splitting a story into a series of e-books and offering the first e-book as free content, a strategy that the platform actively recommends to authors.[16] The number of weeks that have passed since an e-book has been published has a negative effect on whether an e-book is offered as free content in columns (2) and (3) ($p < 0.01$), which is consistent with authors over time building up reputations outside of what is observed through the number of

smaller since the Linear Probability model is estimated on all observations, whereas the Conditional Logit model only uses observations where the dependent variable varies within groups (Beck, 2018), in this case within e-books.

[16] See the Smashword's Book Marketing Guide, p.50: https://www.smashwords.com/books/view/305 (Accessed April 13, 2018)

Table 5 Means of Ratings for Free and Paid E-books

Sample: All E-books	μ	σ	N
Ratings given for paid e-books	4.592	0.7442	51,805
Ratings given for free e-books	4.415	0.9394	47,574
Difference	0.177	-0.1952	
p-value (H_0 : Equal μ, σ)	< 0.001	< 0.001	

Note: Summary statistics in this table are calculated pooling all ratings either into ratings for free or paid versions of e-books. The data contains the information on the review level whether a review was given for a free or paid version.

reviews on the *Smashwords* platform. In column (4), the coefficient on the previous average rating of an e-book is positive ($p < 0.01$), indicating that an average rating that is better by one star increases the probability that an e-book is offered as free content by 7 percent. This result is consistent with authors being more likely to offer e-books as free content that are perceived to be of higher quality, as these e-books are more likely to earn goods reviews, which would help to establish a better reputation. Li et al. (2016) similarly find that a better previous rating makes it more likely that a seller offers a rebate to induce buyers to provide feedback, which is also the theoretical prediction made by the Economics literature on advertising (Nelson, 1974; Kihlstrom and Riordan, 1984; Milgrom and Roberts, 1986).

Table 15 in the Online Appendix shows as a robustness check results where the sample size is restricted to only e-books that were observed in all four waves that the data was collected creating a balanced panel. This restriction has only a negligible impact on the results both in terms of the size of the coefficients and their statistical example.

5.2. Analysis on Review Level

I continue with analyses on the review and rating level, testing whether free content receives worse reviews (Hypothesis 3) and ratings with a higher standard deviation (Hypothesis 4). In the case of Hypothesis 3, I analyze not only whether ratings are worse for a free e-book but also whether written reviews are shorter and have a worse sentiment.

5.2.1. Free Content Receives Worse Ratings. According to Hypothesis 3, content should receive worse reviews when it is offered for free as compared to when it is offered at a positive price. The right panel of figure 1 shows the distribution of ratings for free and paid e-books. Free e-books gather fewer five-star ratings but more one- to four-star ratings compared to paid e-books. This results (see table 5) in a lower average rating of 4.415 stars for free e-books while paid-for e-books received an average rating of 4.592 stars (stat. different with $p < 0.001$). The observed difference in ratings, however, might be due to lower quality e-books being more likely to be offered as free content. To account for such a selection effect, I take advantage of the fact that *Smashwords* provides for each review the information on its website whether the review was given for a free e-book (see figure 3 in the Online Appendix). Therefore, I can estimate the following regression model:

$$\begin{aligned} Rating_{ijt} = &\beta_1 FreeBook_{ijt} \\ &+ x'_{ijt}\boldsymbol{\beta} + b_i + \tau_t + \epsilon_{ijt}, \end{aligned} \tag{8}$$

on the level of rating j given for e-book i in time period t, where $FreeBook_{ijt}$ is a dummy variable indicating whether rating j was given to a free version of e-book i, x'_{ijt} are controls on the level of a rating and ϵ_{ijt} is the error term. As in the previous analyses, I include e-book level fixed effects b_i and time period (month-year when a review was given) fixed effects τ_t. Hence, the coefficients are estimated only through within e-book variation of the variable $FreeBook_{ijt}$. Standard errors ϵ_{ijt} are clustered on the e-book level.

Table 6 shows the results of a series of regression models. Consistent with Hypothesis 3, whether a review is provided for an e-book that is offered as free content has a negative and statistically significant ($p < 0.05$ in columns (1) - (3) and $p < 0.1$ in column (4)) impact on the rating. The size of the effect indicates that an e-book receives ratings that are 0.05 to 0.06 stars lower when it is offered as free content. As the overall standard deviation of ratings is 0.85, this corresponds to a change between 6 to 7 percent of the standard deviation of ratings. The size of the effect does not significantly change when controlling for the number of weeks that have passed since an e-book has been published in column (2), the number of previous reviews of the same e-book in column (3), and the mean of previous ratings and the previous rating in column (4).

The estimated coefficients of the control variables are mostly as expected: The number of weeks that have passed since an e-book has been published and the number of previous reviews have a negative effect on the rating ($p < 0.01$), which is consistent with Li and Hitt (2008), who show that ratings for products decline over time as consumers with a higher preference adopt and review a product earlier in its life-cycle. The negative impact of the mean of previous ratings of an e-book (column 4, $p < 0.01$) on the focal rating is consistent either with higher ratings attracting reviewers with a lower average preference or a higher rating raising consumers' expectations which are then more likely to be not full-filled. Cabral and Li (2015) similarly find that sellers with a better review score are more likely to receive negative feedback and suggest the same explanation. The previous rating of an e-book (column 4, $p < 0.01$) has a positive impact on the focal rating, which might be indicative of herding (Banerjee, 1992; Duan et al., 2009; Lee et al., 2015).

Taken together, these results confirm Hypothesis 3. In section 5.3, I will provide evidence that the negative effect can be explained by a selection effect whereby consumers with a lower preference are more likely to acquire an e-book when it is offered as free content.

Table 6 Impact of Offering E-Book as Free Content on Ratings

	Dependent variable:			
	Rating (1-5 Stars)			
	(1)	(2)	(3)	(4)
Free e-book reviewed (Yes=1)	-0.062^{**}	-0.057^{**}	-0.056^{**}	-0.051^{*}
	(0.024)	(0.024)	(0.024)	(0.027)
Log(N Weeks reviewed after published)		-0.055^{***}	-0.049^{***}	-0.057^{***}
		(0.006)	(0.007)	(0.009)
Log(N previous reviews)			-0.015^{*}	-0.048^{***}
			(0.008)	(0.014)
Previous rating				0.071^{***}
				(0.013)
Average previous ratings				-0.767^{***}
				(0.039)
Fixed effects e-book	*Yes*	*Yes*	*Yes*	*Yes*
Fixed effects month-year	*Yes*	*Yes*	*Yes*	*Yes*
Number e-books	42,496	42,496	42,496	18,094
Observations	99,379	99,379	99,379	57,173

*p<0.1; **p<0.05; ***p<0.01
Clustered standard errors in parentheses

Notes: This table shows linear regression results with fixed effects. Standard errors are clustered on the e-book level. To all variables where the logarithm is taken one is added to account for observations that are zero. Column (5) includes only ratings where previous ratings are available for the same e-book.

As a robustness check, in table 16 in the Online Appendix regression results are shown where only those e-books are included where reviews for both a free and a paid-for version are observed. This reduces the sample size to $13,899$ reviews from $2,059$ e-books. All coefficients and their statistical significance remain virtually unchanged.

5.2.2. Free Content Receives Shorter Written Reviews with Worse Sentiment. Additional to assigning a rating, readers on *Smashwords* have the option to leave a written review that is published together with the rating (see figure 3). In this subsection, I examine whether e-books that are offered as free content receive shorter reviews that are written in a worse sentiment. If free content attracts a larger number of readers with a lower average preference, it is plausible that these readers are also less enthusiastic about the content and as a result write shorter reviews with a less positive sentiment. This might have an additional impact on the success of an e-book, as textual reviews contain additional information that is not captured in numerical ratings and which therefore might be relevant for consumers' purchasing decisions (Archak et al., 2011; Ghose and Ipeirotis, 2011).

As in the previous section, I run regression models on the review level including e-book and time-period fixed effects. Instead of the rating as in section 5.2.1, the dependent variable in these regressions is either the length of a review in terms of the number of characters or the sentiment of a review. To measure the sentiment of reviews, I estimate a text-based sentiment model using reviews of e-books for which either only free or only paid versions have been reviewed as a training set or "ground truth." These reviews can be excluded from the final regression model, as due to the e-book level fixed effects e-books with no variation in the independent variable do not contribute towards estimating the coefficient on the variable. After excluding common stop words from the reviews such as "and", "or", and "if" and reducing all words to their stems, I train a machine learning model on the training set that uses the 4000 most common one-, two- and three-word phrases (uni-, bi- and trigrams in the language of computational linguistics) occurring in reviews to predict the star rating associated with each written review. The model is based on a linear regression model using the Lasso-method to prevent overfitting (Tibshirani, 1996). The model estimates a coefficient for each of the 4000 phrases that measure the impact that the occurrence of a phrase in a written review has on the star rating associated with the review. Based on the estimated coefficients, for each written review a sentiment score can be computed by adding up the coefficients of those phrases that occur in the review. This approach using Lasso-regression to estimate sentiment is the same as in Gentzkow and Shapiro (2010). Table 17 in the Online Appendix shows the twenty phrases whose presence is estimated to have either the most positive or the most negative impact on the prediction of a rating. Phrases such as "excellent" and "cool book" are more likely to be associated with high ratings whereas phrases such as "boring", "annoying" and "proofread" are associated with low ratings.

Table 7 shows the regression results. Columns (1) and (2) show the results of the regression models with the logarithm of the number of characters in a written review as the dependent variable. Columns (3) and (4) show the results from the regression models with the sentiment of a review as the dependent variable. The regression models in columns (3) and (4) only include reviews for e-books where the variable *Free e-book rated* varies within e-books as these reviews were not used in estimating the sentiment model. E-book and time-period fixed effects are included in all models in the same manner as in the previous regression models.

The coefficient on *Free e-book rated* in columns (1) and (2) is negative ($p < 0.01$), confirming that an e-book receives shorter written reviews when it is offered as free content. The coefficient on *Free e-book rated* in column (3) is negative ($p < 0.01$) also confirming that an e-book receives written reviews that have a worse sentiment when it is offered as free content. This effect in itself however is not surprising, as the sentiment score is strongly correlated with ratings by construction. To examine whether there is an effect of offering an e-book as free content on the sentiment additional to what would be predicted just by the correlation between sentiment and ratings, in column (4) I include the actual rating associated with a review as an additional independent variable into the regression model. The coefficient on *Free e-book rated* remains negative and statistically significant (p-value < 0.01), which implies that an e-book receives both lower star-ratings and reviews written in a worse sentiment when it is offered as free content.

Table 7 Impact of Offering E-Book as Free Content on Length and Sentiment of Review

	Dependent variable:			
	Log(N characters in review + 1)		Sentiment	
	(1)	(2)	(3)	(4)
Free e-book rated (Yes=1)	−0.098***	−0.095**	−0.080***	−0.065***
	(0.036)	(0.041)	(0.018)	(0.016)
Actual rating				0.350***
				(0.014)
Log(N weeks rated after published)	−0.007	−0.015	−0.021*	0.004
	(0.010)	(0.012)	(0.011)	(0.009)
Log(N previous ratings)	0.011	0.023	−0.057***	−0.029**
	(0.013)	(0.017)	(0.014)	(0.013)
Previous rating		−0.019	0.010	−0.013
		(0.014)	(0.015)	(0.012)
Mean previous rating		0.051	−0.190***	0.053
		(0.037)	(0.041)	(0.036)
Fixed effects e-book	*Yes*	*Yes*	*Yes*	*Yes*
Fixed effects month-year	*Yes*	*Yes*	*Yes*	*Yes*
Number e-books	42,496	18,094	2,059	2,059
Observations	99,379	57,173	11,856	11,856

*p<0.1; **p<0.05; ***p<0.01
Clustered standard errors in parentheses

Note: This table shows linear regression results with fixed effects. Standard errors are clustered on the e-book level. Sentiment is estimated via Lasso-regression using reviews as a training set excluded from the regression model. To all variables where the logarithm is taken one is added to account for observations that are zero.

5.2.3. Free Content Receives More Dispersed Ratings. In this section, I examine whether the dispersion of ratings of an e-book is larger when it is offered as free content (Hypothesis 4).

Looking at the distribution of ratings in the right panel of figure 1 and the summary statistics of free and paid e-books in table 5 already indicates that ratings for free e-books might have a higher dispersion. To test this within a regression framework, I use the following approach: After demeaning each rating with the overall average rating given to the same e-book, I use the absolute value of each demeaned rating as the dependent variable in my regression models. This newly created variable measures the absolute deviation of a given rating from the average rating of the same e-book.[17] Using this variable, I estimate the following regression equation:

$$|Rating_{ij} - \overline{Rating}_i| = \beta_0 + \beta_1 FreeEbook_{ij} + x'_{ij}\boldsymbol{\beta} + b_i + \tau_t + \epsilon_{i,j}, \tag{9}$$

where $Rating_{ij}$ is the rating j given to e-book i, $\overline{Rating}_i$ is the average rating of e-book i, $FreeEbook_{ij}$ indicates whether rating j was given to a free e-book, x_{ij} are time-varying controls, b_i and τ_t are

[17] For example, if an e-book has received three ratings $1, 3, 5$; the average rating is 3 and the demeaned ratings are $-2, 0, 2$. The absolute values of these demeaned ratings are $2, 0, 2$.

Table 8 Impact of Offering E-book as Free Content on Dispersion of Ratings

	Dependent variable:			
	Abs(rating - mean rating of e-book)			
	(1)	(2)	(3)	(4)
Free e-book rated (Yes=1)	0.042***	0.040***	0.040***	0.051***
	(0.012)	(0.012)	(0.012)	(0.015)
Log(N weeks rated after published)		0.016***	0.015***	0.030***
		(0.003)	(0.003)	(0.005)
Log(N previous ratings)			0.002	0.011
			(0.004)	(0.007)
Previous rating				−0.012*
				(0.007)
Mean previous rating				0.203***
				(0.021)
Fixed effects e-book	*Yes*	*Yes*	*Yes*	*Yes*
Fixed effects month-year	*Yes*	*Yes*	*Yes*	*Yes*
Number e-books	2,059	2,059	2,059	2,059
Observations	99,379	99,379	99,379	57,173

*p<0.1; **p<0.05; ***p<0.01
Clustered standard errors in parentheses

Note: This table shows OLS regression results with fixed effects. The dependent variable is the absolute value of the deviation of a given rating from the average rating of an e-book. Standard errors are clustered on the e-book level. To all variables where the logarithm is taken one is added to account for observations that are zero. Column (4) includes only ratings where a previous rating is observed for the same e-book.

e-book and time period fixed effects, and ϵ_{it} is the error term. The e-book level fixed effects b_i pick up the average deviation from the mean rating for each e-book, such that the coefficient β_1 is only identified trough within e-book changes in the variable $FreeEbookRated_{ij}$.

Table 8 shows the results of the estimated regression models. The coefficient on $FreeEbookRated_{ij}$ is positive and statistically significant in all estimated regression models (p-values < 0.01). The size of the coefficient implies that an e-book that is offered as free content receives ratings that are by 0.04 to 0.05 units more dispersed, which is consistent with Hypothesis 4. As the average standard deviation of ratings is 0.85, the effect size implies an increase in the dispersion of ratings by about 5 to 6 percent.

The estimated coefficients on the control variables imply that i) the dispersion of ratings increases over time, which might be explained by readers with more diverse tastes buying e-books in later stages of an e-book's life-cycle, ii) that the dispersion of ratings increases with a higher mean rating, which similarly might be explained by a higher mean rating attracting readers with more diverse tastes, and iii) that the dispersion in ratings decreases with higher previous ratings, which could be explained by herding of ratings being more likely for higher previous ratings.

5.3. Selection into Reviewing Free Content

In this section, I provide additional evidence in support of selection driving the negative effect of offering content for free on online reviews. To do so, I compute a set of variables for each reviewer, showing that differences in these variables between those reviewers who provide reviews for free e-books and those who provide reviews for paid e-books can account for the observed difference in ratings between free and paid e-books. Thereby, I rule out alternative explanations such as price directly influencing the utility of consumers (Shiv et al., 2005; Shampanier et al., 2007) or their propensity to provide good reviews (Li and Hitt, 2010) as driving the main effect.

To compute several variables capturing the characteristics of reviewers, I use the fact that users who sign up to the *Smashwords* platform have to choose a unique username, which is displayed next to each review of an e-book. Using this information, I compute the following variables for each reviewer: the number of previous reviews, the likely gender based on a reviewer's first name,[18] the number of previously reviewed free e-books, and the mean rating of previous reviews. I term these variables *Reviewers' characteristics* as they are not directly dependent on the e-book a reviewer is currently reviewing. Additionally, I compute for each review based on the e-book and the previously reviewed e-books of the same reviewer the following variables: whether a reviewer has previously reviewed an e-book by the same author, an e-book of the same genre or sub-genre, and whether the currently reviewed e-book is from the reviewer's favorite genre or sub-genre, defined as the genre or sub-genre that a reviewer has previously written the most reviews for. I term these variables *Match characteristics* as they depend on the e-book a reviewer is currently reviewing and are therefore indicative of whether a reviewer's personal taste matches the reviewed e-book.

Table 9 shows summary statistics computed for each variable, splitting the data into reviews for free and paid e-books. There are notable differences in the variables between reviews of free and paid e-books. For example, free e-books are reviewed by reviewers who on average have previously reviewed 48.9 other e-books, of which 45.2 were free e-books, while non-free e-books are reviewed by reviewers who have previously reviewed only 7.5 other e-books, of which 4.0 were free e-books. Additionally, free e-books are more likely to be reviewed by male reviewers, by reviewers who have not previously reviewed the same author, by reviewers who have previously reviewed e-books from the same genre or sub-genre, and by readers who have a different favorite genre or sub-genre ($p < 0.01$ in all cases). Most importantly, reviewers of free e-books have assigned on average lower ratings to previously reviewed e-books than the reviewers of paid e-books ($p < 0.01$). Overall, these differences show that there is significant selection in terms of which type of reader acquires and reviews free or paid e-books.

To show that the computed variables on the reviewer level also impact ratings and therefore differences in these variables can explain differences in ratings, I include the computed variables into

[18] I use the `R` package `gendeR` that predicts gender based on data from various historical databases such as the U.S. Census or the U.S. Social Security Administration database (Mullen, 2018). For reviewers who do not post their first name on the platform or whose name is not found in the `gendeR` database, I set the probability of being female equal to the mean in the data.

Table 9 Summary of Reviewers' Characteristics

		Reviews free e-books (N=51,805)	Reviews paid e-books (N=47,574)	
		Mean	Mean	Diff. means
Reviewers' characteristics	Number previous reviews by same reviewer	48.86	7.516	41.35***
	Gender reviewer (1=female)	0.560	0.631	−0.0711***
	Number previously reviewed free e-books by same reviewer	45.16	3.96	41.2***
	Mean of previous ratings by same reviewer	4.34	4.48	−0.136***
Match characteristics	Reviewer previously reviewed same author (1=yes)	0.121	0.163	−0.042***
	Reviewer previously reviewed same genre (1=yes)	0.578	0.404	0.1734***
	Reviewer previously reviewed same sub-genre (1=yes)	0.422	0.266	0.156***
	From reviewer's favorite genre (1=yes)	0.951	0.973	−0.022***
	From reviewer's favorite sub-genre (1=yes)	0.632	0.760	−0.128***

*p<0.1; **p<0.05; ***p<0.01

Note: This table shows average characteristics of reviewers of reviews given to free and to non-free e-books.

the regression model where a single rating is the dependent variable. Table 10 shows the results of a series of regression models where time-constant heterogeneity between e-books is controlled for by e-book level fixed effects in the same manner as in the regression models in the previous sections.

The results of the full model in column (3), which includes both reviewers' and match characteristics and all variables except the previous rating and the mean of previous ratings of the same e-book, show that female reviewers, reviewers with a higher mean rating of previous reviews, and reviewers who have previously reviewed the same author assign on average higher ratings (p-value at least < 0.1 in all cases). Generally, these results are consistent with readers who have a higher preference for an e-book to also give higher ratings. Notably, the main effect of whether a review was given to a free version of an e-book decreases by half in size and the remaining effect is no longer statistically significant ($p-$value $= 0.49$). Although it is not clear whether the effect is actually zero or merely not statistically different from zero, this result still provides evidence that the negative effect on ratings is mainly explained by reviewers with different characteristics reviewing an e-book when it is offered as free content.

To summarize the net effect of selection based on reviewers' characteristics on ratings, I continue with an analysis where I predict for each observed rating in the dataset the rating the reviewer would have given based on her characteristics. I then use the predicted rating as the dependent variable in a regression model. This allows me to estimate whether there is a difference in predicted ratings between free and paid e-books controlling for the same set of variables and e-book level fixed effects as in the previous analyses. A difference in predicted ratings between free and paid e-books would be evidence in support of selection of reviewers explaining the difference in actual ratings.

Figure 5 in the Online Appendix shows density estimates of predicted ratings splitting the dataset into ratings for free and paid e-books. The prediction is based on a linear regression model that

Table 10 Impact of Reviewer's Characteristics on Ratings

	Dependent variable:			
	Rating (1-5 Stars)			
	(1)	(2)	(3)	(4)
Free e-book rated (Yes=1)	-0.056^{**}	-0.032	-0.025	-0.023
	(0.024)	(0.037)	(0.037)	(0.040)
Log(N Weeks rated after published)	-0.049^{***}	-0.024^{**}	-0.022^{**}	-0.025^{**}
	(0.007)	(0.010)	(0.010)	(0.012)
Log(N previous reviews same e-book)	-0.015^{*}	-0.037^{***}	-0.035^{***}	-0.079^{***}
	(0.008)	(0.013)	(0.013)	(0.021)
Previous rating				0.051^{***}
				(0.016)
Mean previous rating				-0.634^{***}
				(0.053)
Reviewers' Characteristics				
Log(N previously reviewed e-books same reviewer)		-0.042^{***}	-0.045^{***}	-0.024
		(0.014)	(0.014)	(0.015)
Gender reviewer (1=female)		0.044^{**}	0.046^{**}	0.061^{***}
		(0.018)	(0.018)	(0.019)
Log(N previously reviewed free e-books)		0.023	0.023^{*}	0.006
		(0.014)	(0.014)	(0.015)
Mean of previous ratings by same reviewer		0.488^{***}	0.482^{***}	0.423^{***}
		(0.016)	(0.016)	(0.016)
Match Characteristics				
Reviewer previously reviewed same author (1=yes)			0.120^{***}	0.084^{***}
			(0.019)	(0.020)
Reviewer previously reviewed same genre (1=yes)			-0.038	-0.016
			(0.025)	(0.025)
Reviewer previously reviewed same sub-genre (1=yes)			0.001	-0.00000
			(0.019)	(0.019)
From reviewer's favorite genre (1=yes)			0.008	-0.014
			(0.063)	(0.063)
From reviewer's favorite sub-genre (1=yes)			-0.005	0.002
			(0.017)	(0.018)
Fixed effects e-book	*Yes*	*Yes*	*Yes*	*Yes*
Fixed effects month-year	*Yes*	*Yes*	*Yes*	*Yes*
Number e-books	42,496	29,582	29,582	12,982
Observations	99,379	55,352	55,352	30,534

*p<0.1; **p<0.05; ***p<0.01
Clustered standard errors in parentheses

Note: This table shows linear regression results with fixed effects. To all variables where the logarithm is taken one is added to account for observations that are zero. Standard errors are clustered on the e-book level.

includes all nine reviewers' and match characteristics variables that are summarized in table 9 as independent variables but no fixed effects and none of the independent variables that are later included in the regression model. All nine variables have a statistically significant impact on ratings and the regression model has an R-Squared of 0.275 showing that it has significant explanatory power. Predicted ratings for free e-books have a lower mean than for paid e-books (4.34 vs. 4.49; p-value < 0.001) which is also visible in the distributions of predicted ratings shown in figure 5.

In table 11, regression results are shown with the predicted rating as the dependent variable. In column (1), only e-book level fixed effects are included, whereas in column (2) month-year fixed effects are also included. In column (3), standard errors are additionally clustered on the e-book level.

Table 11 Impact of Offering E-Book as Free Content on Predicted Rating

	Dependent variable:		
	Predicted Rating (1-5 Stars)		
	(1)	(2)	(3)
Free e-book reviewed (Yes=1)	-0.034^{**}	-0.033^{**}	-0.033
	(0.016)	(0.017)	(0.024)
Log(N weeks reviewed after published)	-0.043^{***}	-0.041^{***}	-0.041^{***}
	(0.004)	(0.005)	(0.007)
Log(N previous reviews)	0.001	0.003	0.003
	(0.008)	(0.008)	(0.011)
Previous rating	0.020^{***}	0.018^{***}	0.018^{*}
	(0.006)	(0.006)	(0.009)
Mean previous rating	-0.217^{***}	-0.214^{***}	-0.214^{***}
	(0.017)	(0.017)	(0.027)
Fixed Effects E-Book	*Yes*	*Yes*	*Yes*
Fixed Effects Month - Year Reviewed	*No*	*Yes*	*Yes*
Clustered Standard Errors	*No*	*No*	*Yes*
Number e-books	12,982	12,982	12,982
Observations	30,534	30,534	30,534

*p<0.1; **p<0.05; ***p<0.01
Standard errors in parentheses

Note: This table shows linear regression results with fixed effects. To all variables where the logarithm is taken one is added to account for observations that are zero. In the case of clustered standard errors, they are clustered on the e-book level.

The size of the coefficients on the covariate *Free E-Book* have the same size in all three regression models, indicating that an e-book receives a predicted rating that is by 0.03 stars lower when it is offered for free. The coefficient is statistically significant in columns (1) and (2), where fixed effects are included, but not in column (3) where standard errors are clustered on the e-book level and therefore might be too large to identify an effect ($p = 0.18$). The size of the coefficient has a similar magnitude as in the previous analyses where the actual rating was the dependent variable, which suggests selection as the main explanation for the observed worse ratings for free e-books.

6. Discussion and Limitations

In this section, I discuss the results of this work in relation to some selected previous literature and highlight some limitations and directions for future work.

As explained in the introduction, the impact of offering content for free on online reviews could potentially occur *within-consumers* or *across-consumers*. Overall, I find that offering content for free has a negative impact on ratings, and the analysis in section 5.3 lends support to the hypothesis that the observed effect is explained by an *across-consumer* selection effect. After controlling for observed characteristics of reviewers, the remaining effect of offering content for free is statistically not different from zero. However, the coefficient is still negative, so it is possible that there is still a remaining negative effect that is not statistically distinguishable from zero due to a lack of statistical power. Therefore, the cautious interpretation of my results is that the *across-consumer* selection effect based on observable reviewer characteristics dominates any possible within-consumer effects in my empirical setting. However, within-consumer effects might still occur, which could be explored by future research either using more detailed observational data or experimental methodologies.

A limitation of the data used in this study is that it only allows for distinguishing between whether a review was given for content offered for free or content offered at a price. Therefore, it is not possible to draw strong conclusions on whether the established effects on ratings are driven by free being merely a lower price or whether there is anything "special" about a zero price (Shampanier et al., 2007). However, the theoretically motivated selection of consumers induced by "free" being a lower price is sufficient to explain the empirical patterns observed in the data. Future research could therefore examine the extent to which a low but non-zero price induces a similar negative selection effect of consumers on online reviews and reputation.

A study that is particularly close to this work is Lin et al. (2019) who study the impact of free sample provision on online reputation in the context of physical goods sold on an e-commerce platform. They show that free samples can induce more positive online reviews due to consumers reciprocating a free product by providing a better review. In the context of free content provision, this study does not find evidence for such a positive reciprocity effect. There are three possible explanations: First, providing free content does not induce a direct cost on the content provider, as digital content is produced and distributed at zero marginal costs. Therefore, consumers might be less likely to reciprocate. Second, on the e-book platform studied in this work about 18 % of e-books are offered as free content over the time-frame that the data were collected. Therefore, consumers might not perceive receiving an e-book as free content as anything special and as a consequence do not reciprocate. This might be different in contexts such as the e-commerce platform that is studied by Lin et al. (2019). Exploring such boundary conditions and contextual factors that might trigger reciprocal behavior is an interesting avenue for future research. Third, in order for free to induce a negative selection effect on online reviews based on consumer preferences, it is necessary that consumer preferences are sufficiently heterogeneous. As in the setting of e-books studied in this work consumer heterogeneity in terms of tastes is likely to be particularly high, it is especially likely to find an empirically relevant selection effect. Lin et al. (2019) on the other hand conduct their study in the context of mostly physical products that include several categories of consumer products such as household items and home appliances for which consumer preferences are likely to be substantially less heterogeneous compared to experience goods such as e-books.This is also a possible explanation why the results of this study are in contrast to those of Li and Hitt (2010) who find a positive effect of price on online reviews using data of online reviews for digital cameras and Li (2016) in the context of restaurants. To support this argument that the negative selection effect is larger for product categories that exhibit larger degrees of consumer heterogeneity, I have conducted a supplementary analysis exploiting the fact that different genres of e-books exhibit different degrees of consumer heterogeneity as measured by the within genre average standard deviation of ratings. Table 18 in the Online Appendix show the standard deviations for the different genres of e-books in the dataset. Figure 4 and the regression analysis in table 19 in the Online Appendix show that the negative effect of free content provision increases with the within-genre variation of reviews. This supports the conclusion that sufficiently large consumer heterogeneity is a prerequisite for a negative selection effect on ratings to occur.

7. Conclusion

As online reputation systems become more prevalent throughout the digital economy, it becomes more important for sellers and firms to actively manage their online reputation. This can be especially a challenge for entrants into online marketplaces without an established online reputation. In this paper, I look at how content providers provide free products in the form of free digital content as a strategy to build up a reputation. My findings show there is a trade-off involved: Although free content receives more online reviews, these reviews are worse and their associated ratings have a higher standard deviation compared to the case where the same content is offered at a price. These effects are explained by a selection effect, whereby a zero price attracts more readers who have on average a lower preference for the particular content and who therefore on average provide worse reviews.

While the main contribution of this paper is to the literature on online reputation and free content provision, the results can be interpreted more broadly. In one view, an online reputation is just a more formal and systematic way of measuring the general reputation of a firm (Cabral, 2012b, p.343). Interpreting reputation in this way opens up the question of whether there is a similar selection-induced price-effect for the general reputation of a firm. Although this goes beyond the scope of this paper, the question is whether firms can exploit the price-effect on reputation strategically by setting a high price to induce more favorable word-of-mouth and thus building up a good reputation. Whether such a strategy would also hold up in an equilibrium where consumers have rational expectations of firms' strategies or whether it requires some form of bounded rationality of consumers is not clear. This is a question that could be explored by future research.

For managers, this study has the following implications: One is that sellers trying to build a reputation should target free products at consumers who are likely to have a high preference for the sellers' products. For example, in the e-book market, free content should be targeted at readers who have bought previous e-books of the same content provider or e-books from the same genre. This would not only decrease the negative selection effect but content providers might even be able to induce a positive selection effect building up a better reputation. In terms of platform policy, a negative selection effect on reputation implies an additional entry barrier for sellers and firms entering online marketplaces and platforms with asymmetric information. Marketplaces and platforms could help to lower the entry barrier by providing quality certification themselves or by debiasing reputation measures, for example by giving weight to more objective reviewers who have a weaker personal taste for the products they review.

References

Acemoglu, D., Makhdoumi, A., Malekian, A., and Ozdaglar, A. (2017). "Fast and slow learning from reviews." *NBER Working Paper No. 24046*.

Aguiar, L., Claussen, J., and Peukert, C. (2018). "Catch me if you can: Effectiveness and consequences of online copyright enforcement." *Information Systems Research*, *29*(3), 656–678.

Ajorlou, A., Jadbabaie, A., and Kakhbod, A. (2016). "Dynamic pricing in social networks: The word-of-mouth effect." *Management Science*, *64*(2), 971–979.

Anderson, C. (2004). "The long tail." *Wired Magazine*, *12*(10), 170–177.

Anderson, C. (2009). *Free: The future of a radical price*. Random House.

Anderson, L. R., and Holt, C. A. (1997). "Information cascades in the laboratory." *American Economic Review*, *87*(3), 847–862.

Archak, N., Ghose, A., and Ipeirotis, P. G. (2011). "Deriving the pricing power of product features by mining consumer reviews." *Management Science*, *57*(8), 1485–1509.

Ba, S., and Pavlou, P. A. (2002). "Evidence of the effect of trust building technology in electronic markets: Price premiums and buyer behavior." *MIS Quarterly*, *26*(3), 243–268.

Bagwell, K. (2007). "The economic analysis of advertising." In R. Schmalensee (Ed.), *Handbook of Industrial Organization*, vol. 3, 1701–1844, Amsterdam: Elsevier.

Banerjee, A. V. (1992). "A simple model of herd behavior." *Quarterly Journal of Economics*, *107*(3), 797–817.

Bapna, R., Ramaprasad, J., and Umyarov, A. (2018). "Monetizing freemium communities: Does paying for premium increase social engagement?" *MIS Quarterly*, *42*(719–735).

Bapna, R., and Umyarov, A. (2015). "Do your online friends make you pay? A randomized field experiment on peer influence in online social networks." *Management Science*, *61*(8), 1902–1920.

Bar-Isaac, H., and Deb, J. (2014a). "(Good and bad) Reputation for a servant of two masters." *American Economic Journal: Microeconomics*, *6*(4), 293–325.

Bar-Isaac, H., and Deb, J. (2014b). "What is a good reputation? Career concerns with heterogeneous audiences." *International Journal of Industrial Organization*, *34*, 44–50.

Bar-Isaac, H., and Tadelis, S. (2008). "Seller reputation." In *Foundations and Trends in Microeconomics*, vol. 4, 273–351, Boston: Now Publishers.

Bawa, K., and Shoemaker, R. (2004). "The Effects of Free Sample Promotions on Incremental Brand Sales." *Marketing Science*, *23*(3), 345–363.

Beck, N. (2018). "Estimating grouped data models with a binary dependent variable and fixed effects: What are the issues." *arXiv preprint arXiv:1809.06505*.

Bergemann, D., and Välimäki, J. (2000). "Experimentation in markets." *Review of Economic Studies*, *67*(2), 213–234.

Bergemann, D., and Välimäki, J. (2006). "Dynamic pricing of new experience goods." *Journal of Political Economy*, *114*(4), 713–743.

Bikhchandani, S., Hirshleifer, D., and Welch, I. (1992). "A theory of fads, fashion, custom, and cultural change as informational cascades." *Journal of Political Economy*, *100*(5), 992–1026.

Bolton, G., Greiner, B., and Ockenfels, A. (2013). "Engineering trust: Reciprocity in the production of reputation information." *Management Science*, *59*(2), 265–285.

Bolton, G. E., Katok, E., and Ockenfels, A. (2004). "How effective are electronic reputation mechanisms? An experimental investigation." *Management Science*, *50*(11), 1587–1602.

Bose, S., Orosel, G., Ottaviani, M., and Vesterlund, L. (2006). "Dynamic monopoly pricing and herding." *RAND Journal of Economics*, *37*(4), 910–928.

Boudreau, K. J. (2018). "Amateurs crowds & professional entrepreneurs as platform complementors." *NBER Working Paper No. 24512*.

Bowker (2014). "Self-publishing in the United States, 2007-2013." *http://media.bowker.com/documents/bowker/selfpublishing/report2013.pdf*, *Accessed on July 12, 2016*.

Brandes, L., Godes, D., and Mayzlin, D. (2018). "Extremity bias in online reviews: A field experiment." *Working Paper*.

Brynjolfsson, E., Hu, Y. J., and Smith, M. (2006). "From niches to riches: Anatomy of the long tail."

Burtch, G., Hong, Y., Bapna, R., and Griskevicius, V. (2017). "Stimulating online reviews by combining financial incentives and social norms." *Management Science*, *64*(5), 2065–2082.

Byers, J. W., Mitzenmacher, M., and Zervas, G. (2012). "The groupon effect on yelp ratings: A root cause analysis." *Proceedings of the 13th ACM conference on electronic commerce*, 248–265.

Cabral, L. (2005). "The Economics of trust and reputation: A primer." *Working Paper*.

Cabral, L. (2012a). "Oligopoly dynamics." *International Journal of Industrial Organization*, *30*(3), 278—282.

Cabral, L. (2012b). "Reputation on the Internet." In M. Peitz, and J. Waldfogel (Eds.), *Oxford Handbook of the Digital Economy*, Oxford: Oxford University Press.

Cabral, L., and Hortacsu, A. (2010). "The dynamics of seller reputation: Evidence from eBay." *Journal of Industrial Economics*, *58*(1), 54–78.

Cabral, L., and Li, L. I. (2015). "A dollar for your thoughts: Feedback-conditional rebates on eBay." *Management Science*, *61*(9), 2052–2063.

Celen, B., and Kariv, S. (2004). "Distinguishing informational cascades from herd behavior in the laboratory." *American Economic Review*, *94*(3), 484–498.

Chaudhari, H., and Byers, J. (2017). "Impact of free app promotion on future sales: A case study on Amazon appstore." *Working Paper available at SSRN*.

Chellappa, R. K., and Shivendu, S. (2005). "Managing piracy: Pricing and sampling strategies for digital experience goods in vertically segmented markets." *Information Systems Research*, *16*(4), 400–417.

Cheng, H. K., and Liu, Y. (2012). "Optimal software free trial strategy: The impact of network externalities and consumer uncertainty." *Information Systems Research*, *23*(2), 488–504.

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

Choi, A. A., Cho, D., Yim, D., Moon, J. Y., and Oh, W. (2019). "When seeing helps believing: The interactive effects of previews and reviews on e-book purchases." *Information Systems Research*, *30*(4), 1164–1183.

Claussen, J., Kretschmer, T., and Mayrhofer, P. (2013). "The effects of rewarding user engagement: The case of Facebook apps." *Information Systems Research*, *24*(1), 186–200.

Cripps, M. W. (2009). *Reputation*. New Palgrave Dictionary of Economics, Basingstoke: Palgrave Macmillan, 2nd edn.

Dellarocas, C. (2003). "The digitization of word of mouth: Promise and challenges of online feedback mechanisms." *Management Science*, *49*(10), 1407–1424.

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

Dellarocas, C., and Wood, C. A. (2008). "The sound of silence in online feedback: Estimating trading risks in the presence of reporting bias." *Management science*, *54*(3), 460–476.

Dellarocas, C., Zhang, X. M., and Awad, N. F. (2007). "Exploring the value of online product reviews in forecasting sales: The case of motion pictures." *Journal of Interactive Marketing*, *21*(4), 23–45.

Dewan, S., and Ramaprasad, J. (2012). "Research note—Music blogging, online sampling, and the long tail." *Information Systems Research*, *23*(3-part-2), 1056–1067.

Dewan, S., and Ramaprasad, J. (2014). "Social media, traditional media, and music sales." *MIS Quarterly*, *38*(1), 101–122.

Duan, W., Gu, B., and Whinston, A. B. (2008). "Do online reviews matter? - An empirical investigation of panel data." *Decision Support Systems*, *45*(4), 1007–1016.

Duan, W., Gu, B., and Whinston, A. B. (2009). "Informational cascades and software adoption on the internet: An empirical investigation." *MIS Quarterly*, *33*(1), 23–48.

Einav, L., Farronato, C., Levin, J., and Sundaresan, N. (2018). "Auctions versus posted prices in online markets." *Journal of Political Economy*, *126*(1), 178–215.

Einav, L., Kuchler, T., Levin, J., and Sundaresan, N. (2015). "Assessing sale strategies in online markets using matched listings." *American Economic Journal: Microeconomics*, *7*(2), 215–247.

Elfenbein, D. W., Fisman, R., and McManus, B. (2012). "Charity as a substitute for reputation: Evidence from an online marketplace." *Review of Economic Studies*, *79*(4), 1441–1468.

Elfenbein, D. W., Fisman, R., and McManus, B. (2015). "Market structure, reputation, and the value of quality certification." *American Economic Journal: Microeconomics*, *7*(4), 83–108.

Forman, C., Ghose, A., and Wiesenfeld, B. (2008). "Examining the relationship between reviews and sales: The role of reviewer identity disclosure in electronic markets." *Information Systems Research*, *19*(3), 291–313.

Fradkin, A., Grewal, E., Holtz, D., and Pearson, M. (2016). "The determinants of online review informativeness: Evidence from field experiments on Airbnb." *Working Paper*.

Friedman, E., Resnick, P., and Sami, R. (2007). "Manipulation-resistant reputation systems." *Algorithmic Game Theory*, 677–697.

Gao, G. G., Greenwood, B. N., Agarwal, R., and McCullough, J. (2015). "Vocal minority and silent majority: How do online ratings reflect population perceptions of quality?" *MIS Quarterly*, *39*(3), 565–589.

Gentzkow, M., and Shapiro, J. M. (2010). "What drives media slant? Evidence from US daily newspapers." *Econometrica*, *78*(1), 35–71.

Ghose, A., and Ipeirotis, P. G. (2011). "Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics." *IEEE Transactions on Knowledge and Data Engineering*, *23*(10), 1498–1512.

Godinho de Matos, M., Ferreira, P., Smith, M. D., and Telang, R. (2016). "Culling the herd: Using real-world randomized experiments to measure social bias with known costly goods." *Management Science*, *62*(9), 2563–2580.

Goes, P. B., Lin, M., and Au Yeung, C.-m. (2014). ""Popularity Effect" in User-Generated Content: Evidence from Online Product Reviews." *Information Systems Research*, *25*(2), 222–238.

Goh, K.-Y., Heng, C.-S., and Lin, Z. (2013). "Social media brand community and consumer behavior: Quantifying the relative impact of user-and marketer-generated content." *Information Systems Research*, *24*(1), 88–107.

Heiman, A., McWilliams, B., Shen, Z., and Zilberman, D. (2001). "Learning and forgetting: Modeling optimal product sampling over time." *Management Science*, *47*(4), 532–546.

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

Johar, M., Kumar, N., and Mookerjee, V. (2012). "Content provision strategies in the presence of content piracy." *Information Systems Research*, *23*(3-part-2), 960–975.

Juan, F., Xin, L., and Zhang, M. X. (2019). "Online product reviews-triggered dynamic pricing: Theory and evidence." *Information Systems Research*, *30*(4), 1107–1123.

Kalish, S. (1983). "Monopolist pricing with dynamic demand and production cost." *Marketing Science*, *2*(2), 135–159.

Kalish, S. (1985). "A new product adoption model with price, advertising, and uncertainty." *Management Science*, *31*(12), 1569–1585.

Khern-am nuai, W., Kannan, K., and Ghasemkhani, H. (2018). "Extrinsic versus intrinsic rewards for contributing reviews in an online platform." *Information Systems Research*, *29*(4), 871–892.

Kihlstrom, R. E., and Riordan, M. H. (1984). "Advertising as a signal." *Journal of Political Economy*, *92*(3), 427–450.

Kretschmer, T., and Peukert, C. (forthcoming). "Video killed the radio star? Online music videos and digital music sales." *Information Systems Research*.

Lambrecht, A., and Misra, K. (2017). "Fee or Free: When Should Firms Charge for Online Content?" *Management Science*, *63*(4), 1150–1165.

Lee, Y.-J., Hosanagar, K., and Tan, Y. (2015). "Do I follow my friends or the crowd? Information cascades in online movie ratings." *Management Science*, *61*(9), 2241–2258.

Lee, Y.-J., and Tan, Y. (2013). "Effects of different types of free trials and ratings in sampling of consumer software: An empirical study." *Journal of Management Information Systems*, *30*(3), 213–246.

Li, L., and Xiao, E. (2014). "Money talks: Rebate mechanisms in reputation system design." *Management Science*, *60*(8), 2054–2072.

Li, L. I. (2010). "Reputation, trust, and rebates: How online auction markets can improve their feedback mechanisms." *Journal of Economics & Management Strategy*, *19*(2), 303–331.

Li, L. I., Tadelis, S., and Zhou, X. (2016). "Buying reputation as a signal of quality: Evidence from an online marketplace." *NBER Working Paper No. 22584*.

Li, X. (2016). "Could deal promotion improve merchants' online reputations? The moderating role of prior reviews." *Journal of Management Information Systems*, *33*(1), 171–201.

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

Li, X., and Hitt, L. M. (2010). "Price effects in online product reviews: An analytical model and empirical analysis." *MIS Quarterly*, *34*(4), 809–831.

Lin, Z., Zhang, Y., and Tan, Y. (2019). "An empirical study of free product sampling and rating bias." *Information Systems Research*, *30*(1), 260–275.

Liu, Y. (2006). "Word of mouth for movies: Its dynamics and impact on box office revenue." *Journal of Marketing*, *70*(3), 74–89.

Luca, M. (2017). "Designing online marketplaces: Trust and reputation mechanisms." *Innovation Policy and the Economy*, *17*(1), 77–93.

Luca, M., and Zervas, G. (2016). "Fake it till you make it: Reputation, competition, and Yelp review fraud." *Management Science*, *62*(12), 3412–3427.

Lucking-Reiley, D., Bryan, D., Prasad, N., and Reeves, D. (2007). "Pennies from eBay: The determinants of price in online auctions." *Journal of Industrial Economics*, *55*(2), 223–233.

MacLeod, W. B. (2007). "Reputations, relationships, and contract enforcement." *Journal of Economic Literature*, 595–628.

Mailath, G. J., and Samuelson, L. (2006). *Repeated games and reputations*. Oxford: Oxford University Press.

Marks, L. J., and Kamins, M. A. (1988). "The use of product sampling and advertising: Effects of sequence of exposure and degree of advertising claim exaggeration on consumers' belief strength, belief confidence, and attitudes." *Journal of Marketing Research*, *25*(3), 266–281.

Mayzlin, D., Dover, Y., and Chevalier, J. (2014). "Promotional reviews: An empirical investigation of online review manipulation." *American Economic Review*, *104*(8), 2421–55.

McFadden, D. L., and Train, K. E. (1996). "Consumers' evaluation of new products: Learning from self and others." *Journal of Political Economy*, *104*(4), 683–703.

Milgrom, P., and Roberts, J. (1986). "Price and advertising signals of product quality." *Journal of Political Economy*, *94*(4), 796–821.

Milgrom, P. R., and Tadelis, S. (2018). "How Artificial Intelligence and Machine Learning can impact market design." *NBER Working Paper No. 24282*.

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

Mullen, L. (2018). *gender: predict gender from names using historical data*. R package version 0.5.2.

Nair, H. (2019). "Diffusion and pricing over the product life cycle." *Working Paper available on SSRN*.

Nelson, P. (1970). "Information and consumer behavior." *Journal of Political Economy*, *78*(2), 311–329.

Nelson, P. (1974). "Advertising as information." *Journal of Political Economy*, *82*(4), 729–754.

Oestreicher-Singer, G., and Zalmanson, L. (2013). "Content or Community? A Digital Business Strategy for Content Providers in the Social Age." *MIS Quarterly*, *37*(2), 591–616.

Oh, H., Animesh, A., and Pinsonneault, A. (2016). "Free versus for-a-fee: The impact of a paywall." *MIS Quarterly*, *40*(1), 31–56.

Pauwels, K., and Weiss, A. (2008). "Moving from free to fee: How online firms market to change their business model successfully." *Journal of Marketing*, *72*(3), 14–31.

Peukert, C., and Reimers, I. (2018). "Digital disintermediation and efficiency in the market for ideas." *Working Paper available on SSRN*.

Proserpio, D., and Zervas, G. (2017). "Online reputation management: Estimating the impact of management responses on consumer reviews." *Marketing Science*, *36*(5), 645–665.

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

Runge, J., Nair, H., and Levav, J. (2019). "Price promotions in 'freemium' settings." *Working Paper available at SSRN*.

Shampanier, K., Mazar, N., and Ariely, D. (2007). "Zero as a special price: The true value of free products." *Marketing science*, *26*(6), 742–757.

Shapiro, C. (1983a). "Optimal pricing of experience goods." *Bell Journal of Economics*, *14*(2), 497–507.

Shapiro, C. (1983b). "Premiums for high quality products as returns to reputations." *Quarterly Journal of Economics*, *98*(4), 659–679.

Shiv, B., Carmon, Z., and Ariely, D. (2005). "Placebo effects of marketing actions: Consumers may get what they pay for." *Journal of Marketing Research*, *42*(4), 383–393.

Smith, M. D., and Telang, R. (2009). "Competing with free: The impact of movie broadcasts on DVD sales and Internet piracy." *MIS Quarterly*, *33*(2), 321–338.

Stenzel, A., Wolf, C., and Schmidt, P. (2019). "Pricing for the stars." *Working Paper available at SSRN*.

Tadelis, S. (2016). "Reputation and feedback systems in online platform markets." *Annual Review of Economics*, *8*, 321–340.

Tibshirani, R. (1996). "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society. Series B (Methodological)*, *58*(1), 267–288.

Tucker, C., and Zhang, J. (2011). "How does popularity information affect choices? A field experiment." *Management Science*, *57*(5), 828–842.

Villas-Boas, J. M. (2004). "Consumer learning, brand loyalty, and competition." *Marketing Science*, *23*(1), 134–145.

Waldfogel, J., and Reimers, I. (2015). "Storming the gatekeepers: Digital disintermediation in the market for books." *Information Economics and Policy*, *31*, 47–58.

Wang, C., Zhang, X., and Hann, I.-H. (2018). "Socially nudged: A quasi-experimental study of friends' social influence in online product ratings." *Information Systems Research*.

Wang, C. A., and Zhang, X. M. (2009). "Sampling of information goods." *Decision Support Systems*, *48*(1), 14–22.

Wibral, M. (2015). "Identity changes and the efficiency of reputation systems." *Experimental Economics*, *18*(3), 408–431.

Xu, J., Forman, C., Kim, J. B., and Van Ittersum, K. (2014). "News media channels: complements or substitutes? Evidence from mobile phone usage." *Journal of Marketing*, *78*(4), 97–112.

Zegners, D. (2016). "Voluntary Disclosure of Product Information: The case of E-book Samples." *Working Paper available on SSRN*.

Zhang, L. (2018). "Intellectual property strategy and the long tail: Evidence from the recorded music industry." *Management Science*, *64*(1), 24–42.

Online Appendix

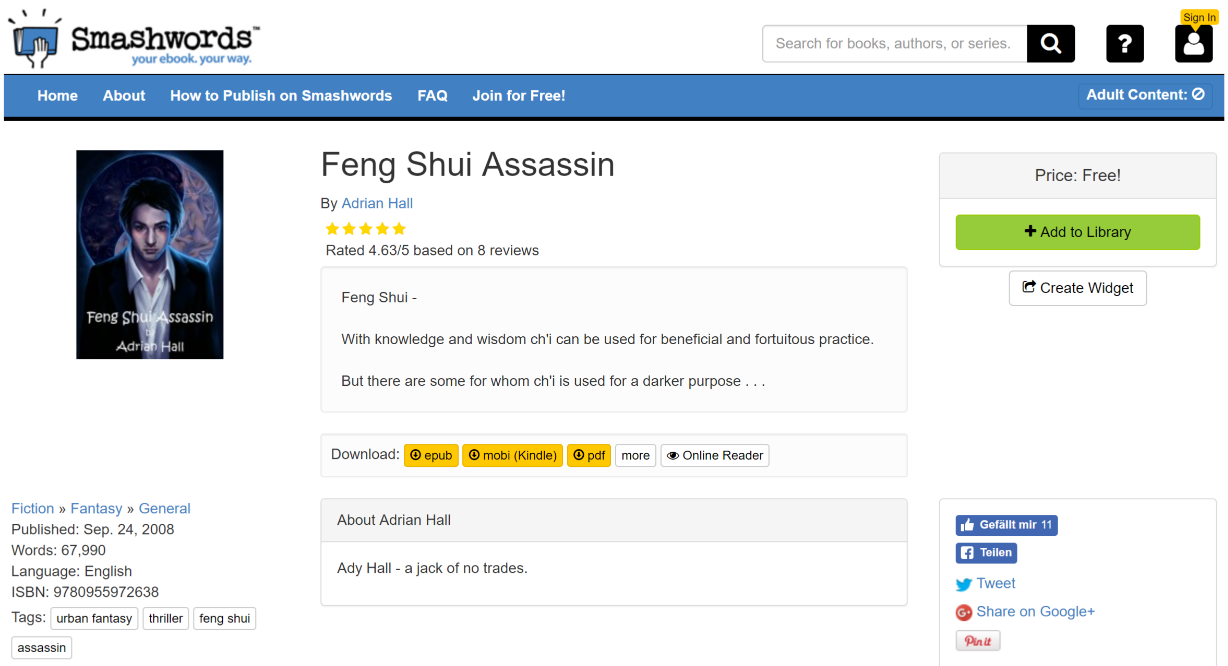


Figure 2 An E-book's Website on Smashwords

Note: This figure shows a screenshot of an e-book's website on *Smashwords*. It shows information such as the name of the author, the e-book's price, the genre of the e-book, the average rating and number of ratings, the language of the e-book etc.

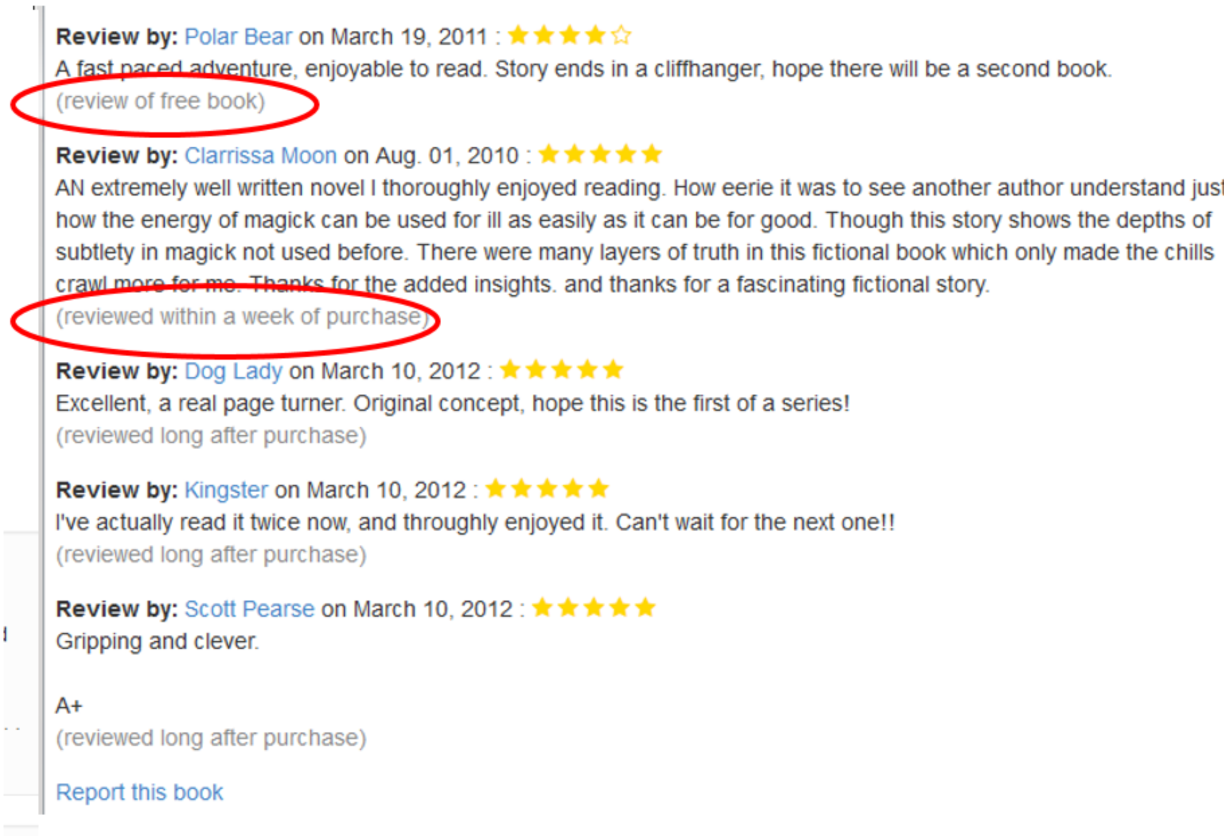


Figure 3 Ratings Posted on Smashwords

Note: This figure shows a screenshot of the ratings posted by readers of an e-book. Additional to the rating assigned (1-5 stars) and a written review, the website shows whether the review was given for a free version ("review of free book") or a paid version ("reviewed within . . . [time]. . . purchase").

Table 12 Impact of Offering E-Book as Free Content on Number of Reviews - Only E-books offered both for Free and at a Price during Time Frame that Data Collected

	Dependent variable:				
	Log(N Reviews + 1, t)				
	(1)	(2)	(3)	(4)	(5)
Free e-book (Yes=1),t-1	0.005***	0.005***	0.004	0.001	0.010**
	(0.002)	(0.002)	(0.003)	(0.003)	(0.004)
Weeks since published, t		0.0002***	0.0002***	0.0001***	0.0001***
		(0.00001)	(0.00001)	(0.00001)	(0.00002)
Log(N e-books same author, t-1)			0.002	−0.005	0.004
			(0.009)	(0.008)	(0.011)
Log(N reviews other e-books same author, t-1)			0.092***	0.031*	0.025*
			(0.016)	(0.017)	(0.013)
Log(N other free e-books same author, t-1)			0.001	0.005	−0.005
			(0.004)	(0.004)	(0.004)
Log(Number reviews, t - 1)				0.288***	0.290***
				(0.024)	(0.052)
Average rating, t-1					0.002
					(0.031)
Fixed effects e-book	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
Fixed effects time-period	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
Observations	25,243	25,243	25,243	25,243	5,719
Number e-books	8,763	8,763	8,763	8,763	2,094

*p<0.1; **p<0.05; ***p<0.01
Clustered standard errors in parentheses

Notes: To all variables where the logarithm is taken one is added to account for observations that are zero. Standard errors are clustered on the e-book level. Column (5) includes only e-books that have at least one rating and were the average previous rating can be computed.

Table 13 Impact of Offering E-Book as Free Content on Number of Reviews - Balanced Panel

	Dependent variable:				
	Log(N Reviews, t)				
	(1)	(2)	(3)	(4)	(5)
Free e-book (Yes=1),t-1	0.012***	0.012***	0.011***	0.008***	0.006*
	(0.002)	(0.002)	(0.002)	(0.002)	(0.003)
Weeks since published, t		0.0001***	0.00005***	0.00004***	0.0001***
		(0.00000)	(0.00000)	(0.00000)	(0.00000)
Log(N e-books same author, t-1)			0.009***	0.006***	0.014***
			(0.002)	(0.001)	(0.004)
Log(N reviews other e-books same author, t-1)			0.068***	0.024***	0.024***
			(0.004)	(0.003)	(0.006)
Log(N other free e-books same author, t-1)			0.0002	0.002*	0.002
			(0.001)	(0.001)	(0.003)
Log(Number reviews, t - 1)				0.370***	0.389***
				(0.008)	(0.015)
Average rating, t-1					−0.003
					(0.008)
Fixed effects e-book	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
Fixed effects time-period	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
Observations	619,701	619,701	619,701	619,701	94,129
Number e-books	206,567	206,567	206,567	206,567	32,248

*p<0.1; **p<0.05; ***p<0.01
Clustered standard errors in parentheses

Notes: Only e-books that were observed in all four time period that data was collected are included. To all variables where the logarithm is taken one is added to account for observations that are zero. Standard errors are clustered on the e-book level. Column (5) includes only e-books that have at least one rating and were the average previous rating can be computed.

Table 14 Drivers of Offering E-book as Free Content - Linear Probability Model

	Dependent variable: Free e-book (Yes=1), t			
	(1)	(2)	(3)	(4)
Log(Number ratings, t-1)	−0.006	−0.006	−0.011***	−0.025**
	(0.004)	(0.004)	(0.004)	(0.011)
Weeks since published, t		0.0001***	0.0001***	0.000
		(0.00000)	(0.00000)	(0.001)
Log(Number books same author,t-1)			0.024***	0.001
			(0.002)	(0.006)
Log(Number ratings other books same author, t - 1)			0.003	−0.007
			(0.002)	(0.006)
Log(Number free e-books same author, t -1)			0.013***	0.041***
			(0.001)	(0.003)
Average rating, t-1				0.010
				(0.010)
Fixed effects e-book	*Yes*	*Yes*	*Yes*	*Yes*
Fixed effects time-period	*Yes*	*Yes*	*Yes*	*Yes*
Number e-books	290,681	290,681	290,681	38,002
Observations	725,060	725,060	725,060	101,625

*p<0.1; **p<0.05; ***p<0.01
Standard errors in parentheses

Note: This table shows linear regression results with fixed effects. To all variables where the logarithm is taken one is added to account for observations that are zero. Column (4) includes only e-books that have at least one rating and were the average previous rating can be computed.

Table 15 Drivers of Offering E-book as Free Content - Logit Model - Balanced Panel

	Dependent variable: Free Book (Yes=1), t			
	(1)	(2)	(3)	(4)
Log(Number Ratings + 1, t-1)	−0.204***	−0.205***	−0.177***	−0.200
	(0.0377)	(0.0377)	(0.0385)	(0.106)
Weeks since published, t		−0.000178***	−0.000015**	−0.00002
		(0.000045)	(0.000049)	(0.00010)
Log(N e-books same author + 1, t-1)			0.0358	−0.0115
			(0.0233)	(0.0519)
Log(N reviews other e-books same author + 1, t-1)			−0.0575*	−0.0236
			(0.0237)	(0.0469)
Log(N other free e-books same author + 1, t-1)			0.0248***	0.0979***
			(0.00630)	(0.0274)
Average rating, t-1				0.0507*
				(0.0786)
Fixed effects e-book	*Yes*	*Yes*	*Yes*	*Yes*
Fixed effects time-period	*Yes*	*Yes*	*Yes*	*Yes*
Number e-books	8,341	8,341	8,341	1,983
Observations	25,023	25,023	25,023	5,921

*p<0.1; **p<0.05; ***p<0.01
Standard errors in parentheses

Note: This table shows results from Conditional Logit regression with fixed effects. Only e-books that were observed in all four time period that data was collected are included. Shown coefficients are computed as average marginal effects. To all variables where the logarithm is taken one is added to account for observations that are zero. Column (5) only includes e-books for which at least one rating is available.

Table 16 Impact of Offering E-Book as Free Content on Ratings - Including Only E-Books With Reviews For Free and Non-free Versions

	Dependent variable:			
	Rating (1-5 Stars)			
	(1)	(2)	(3)	(4)
Free e-book reviewed (Yes=1)	-0.061^{***}	-0.049^{**}	-0.049^{**}	-0.043^{*}
	(0.021)	(0.020)	(0.020)	(0.025)
Log(N Weeks reviewed after published)		-0.075^{***}	-0.053^{***}	-0.072^{***}
		(0.012)	(0.015)	(0.021)
Log(N previous reviews)			-0.045^{***}	-0.081^{***}
			(0.016)	(0.028)
Previous rating				0.066^{***}
				(0.022)
Average previous ratings				-0.694^{***}
				(0.065)
Fixed effects e-book	*Yes*	*Yes*	*Yes*	*Yes*
Fixed effects month-year	*Yes*	*Yes*	*Yes*	*Yes*
Number e-books	2,059	2,059	2,059	2,059
Observations	13,899	13,899	13,899	11,856

*p<0.1; **p<0.05; ***p<0.01
Clustered standard errors in parentheses

Notes: This table shows linear regression results with fixed effects. To all variables where the logarithm is taken one is added to account for observations that are zero. Column (4) includes only reviews where a previous rating for the same e-book is observed. Standard errors are clustered on the e-book level.

Table 17 Words with Largest Estimated Impact on Sentiment

Most Positive Words		*Most Negative Words*	
Word Stem	Coefficient	Word Stem	Coefficient
easi rate	0.389	over good bad	-1.124
bad thing	0.339	wast	-0.728
five stars	0.277	sorry	-0.557
excellent	0.245	boring	-0.448
font write	0.245	potential	-0.418
bad guy	0.240	annoying	-0.405
cool book	0.234	proofread	-0.397
dull moment	0.234	skim	-0.388
five star	0.232	poor	-0.378
grammar errors repetit	0.231	okay	-0.363

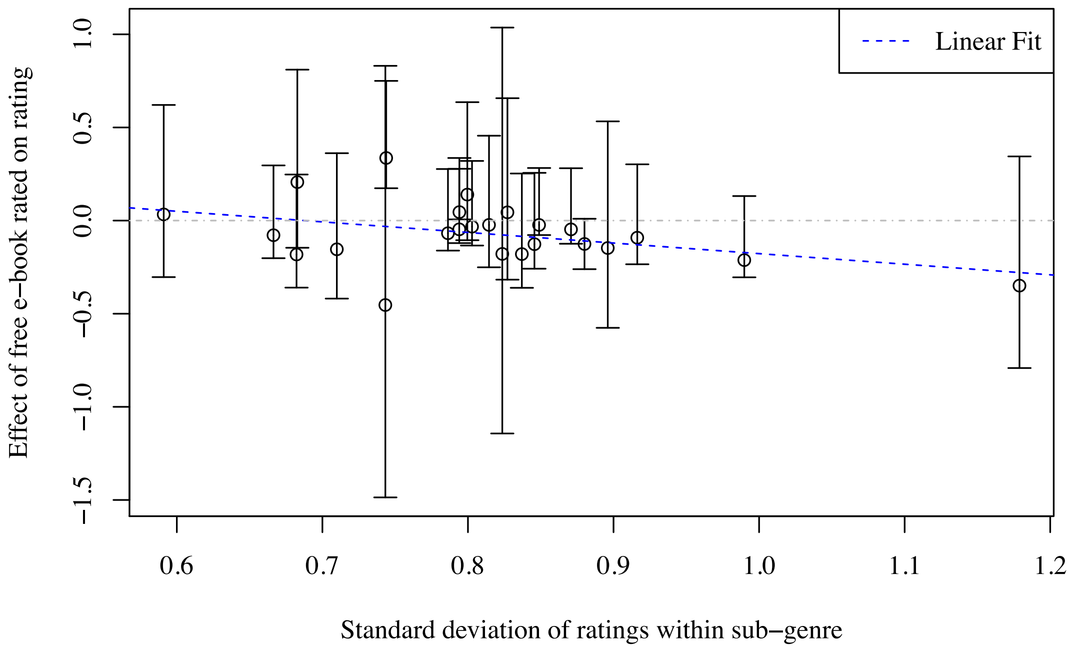
Note: This table shows the coefficients of the 20 uni-, bi- and trigrams which have the most positive (left panel) and most negative (right panel) estimated impact on the predicted sentiment of a review out of the 4000 uni-, bi- and trigrams that were used in the full model. Coefficients are estimated using Lasso-regression on a training set only including reviews for e-books for which either only reviews for free or only paid versions are observed in the data.

Table 18 Standard Deviation of Ratings and Coefficient on Free E-books Across Sub-genres

Sub-genre	Std.Dev.Ratings	Effect Free E-book	N books
Biography	0.591	0.033	1,186
Children's books	0.666	-0.079	2,946
Poetry	0.682	-0.182	1,529
Historical	0.683	0.207	1,845
Travel	0.710	-0.154	443
Self-improvement	0.743	-0.453	745
History	0.744	0.336	345
Fantasy	0.786	-0.068	10,258
Thriller and suspense	0.794	-0.047	4,103
Romance	0.794	0.046	15,326
Anthologies	0.800	0.140	1,589
Mystery and detective	0.803	-0.032	4,514
Business and Economics	0.824	-0.179	853
Gay and lesbian fiction	0.827	0.044	1,964
Literature	0.837	-0.180	2,021
Horror	0.846	-0.127	4,591
Young adult or teen	0.849	-0.023	7,237
Science fiction	0.871	-0.047	7,932
Other (Baseline Category)	0.880	-0.126	13,092
Health, wellbeing, and medicine	0.896	-0.147	734
Humor and comedy	0.916	-0.092	2,218
Erotica	0.990	-0.212	10,539
Religion and Spirituality	1.179	-0.349	1,308

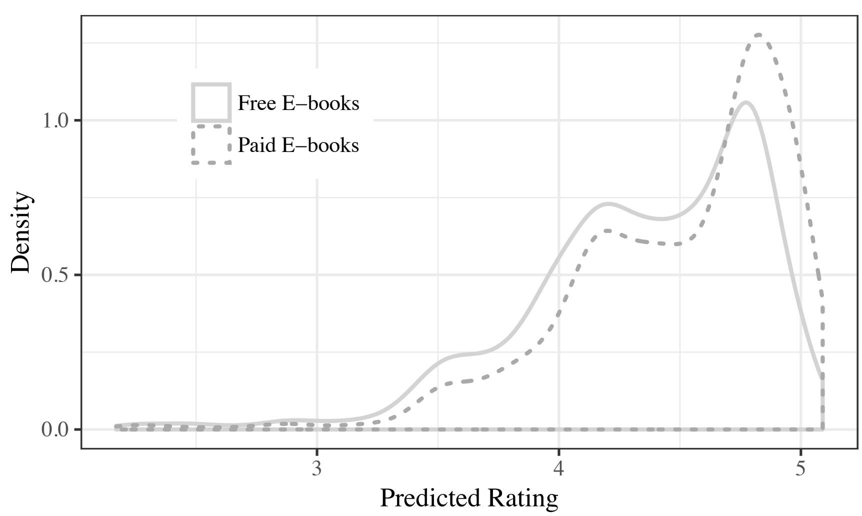
Note: This table shows the standard deviation of ratings and the coefficient of the variable free e-books across different sub-genres of e-books. The coefficients on the variable free e-books are obtained by running the same e-book level fixed effects regression as previously but additionally interacting the free e-book dummy separately with each sub-genre. The values in the "Effect Free E-book" column are obtained by adding the baseline effect (the reference category is "Other") with the coefficients of the interactions with each sub-category dummy.

Figure 4 Effect Size of Free E-book and Within Sub-genre Standard Deviation of Ratings



Note: This graph plots the standard deviations of ratings against the effect size of free e-book across sub-genres shown in table 18. The shown 95% confidence intervals are obtained from the same regression model. The shown fitted line has a negative slope of -0.57 that is statistically different from zero ($p-$value $= 0.049$).

Figure 5 Distribution of Predicted Ratings for Free and Paid E-Books



Note: This plot shows estimated densities using a Gaussian kernel of predicted ratings. Ratings are predicted based on reviewers' and match characteristics shown in table 9.

Table 19 Impact of Offering E-Book as Free Content on Ratings - Interaction with Standard Deviation of Ratings within Sub-genres

	Dependent variable:			
	Rating (1-5 Stars)			
	(1)	(2)	(3)	(4)
Free e-book reviewed (Yes=1)	-0.064^{***}	-0.059^{**}	-0.059^{**}	-0.054^{**}
	(0.024)	(0.024)	(0.024)	(0.027)
Free e-book reviewed X Std. dev. ratings sub-genre	-0.748^{**}	-0.734^{**}	-0.739^{**}	-0.714^{*}
	(0.319)	(0.316)	(0.316)	(0.378)
Log(N Weeks reviewed after published)		-0.055^{***}	-0.049^{***}	-0.057^{***}
		(0.006)	(0.007)	(0.009)
Log(N previous reviews)			-0.015^{*}	-0.048^{***}
			(0.008)	(0.014)
Previous rating				0.071^{***}
				(0.013)
Average previous ratings				-0.766^{***}
				(0.039)
Fixed effects e-book	*Yes*	*Yes*	*Yes*	*Yes*
Fixed effects month-year	*Yes*	*Yes*	*Yes*	*Yes*
Number e-books	42,496	42,496	42,496	18,094
Observations	99,379	99,379	99,379	57,173

*p<0.1; **p<0.05; ***p<0.01
Clustered standard errors in parentheses

Note: This table shows linear regression results with fixed effects. To all variables where the logarithm is taken one is added to account for observations that are zero. Standard errors are clustered on the e-book level.