

MAGKS



**Joint Discussion Paper
Series in Economics**

by the Universities of
**Aachen · Gießen · Göttingen
Kassel · Marburg · Siegen**

ISSN 1867-3678

No. 19-2024

Daniel Lüke

**Tales of Tails: Sales Distribution and the Role of
Retail Channels in the German Book Market**

This paper can be downloaded from

<https://www.uni-marburg.de/en/fb02/research-groups/economics/macroeconomics/research/magks-joint-discussion-papers-in-economics>

Coordination: Bernd Hayo • Philipps-University Marburg
School of Business and Economics • Universitätsstraße 24, D-35032 Marburg
Tel: +49-6421-2823091, Fax: +49-6421-2823088, e-mail: hayo@wiwi.uni-marburg.de

Tales of Tails: Sales Distribution and the Role of Retail Channels in the German Book Market

Daniel Lüke¹

¹University of Giessen, Germany, daniel.lueke@wi.jlug.de

September 2024

Abstract

This paper examines the sales distribution and genre composition of the German book market across different retail channels — e-commerce, chain stores, and independent bookstores — over the period 2011-2018. Utilizing a unique dataset comprising weekly sales data of approximately 50,000 top-selling book titles, the study challenges the economic relevance of the “long tail” effect, which suggests that niche products hold significant market share due to digitalization. Our findings reveal that both online and offline sales are heavily concentrated on a few best-selling titles, with the “long tail” products accounting for a negligible proportion of overall sales. However, a “middle tail” of moderately popular books, more prominent in online sales and independent bookstores, suggests diverse consumption patterns not captured by conventional models focusing on the extremes of sales distributions. We further observe significant differences in the genre composition and attributes of books sold across channels, filling a gap in the existing literature. For example, chain stores exhibit higher concentrations of bestsellers, while independent bookstores show a more diverse array of less recent titles. These findings imply that consumer preferences vary significantly by retail channel, indicating limited substitutability between channels.

JEL classification: L81, L82, Z11

1 Introduction

The impact of digitization on sales distribution and consumption patterns in markets for experience goods, like books, has been widely studied. This paper provides new empirical evidence on two dimensions of the book market: the degree of sales concentration and the differences in sales patterns across retail channels. Based on a comprehensive data set of book sales in Germany, we, first, investigate the distribution of sales, revealing strong concentrations on the most selling books and extremely low sales counts for niche book titles in both, online and offline retail. The extreme extent of the right skewness of the sales distribution challenges the purported economic relevance of this so-called “long tail” in consumption. Second, we explore how consumption differs between retail channels. We find significant differences in genres of the books bought through the respective retail channels. Furthermore, a more disaggregated analysis is presented in which we distinguish not only between e-commerce and brick-and-mortar retail, but additionally distinguish between chain stores and independent book stores within the brick-and-mortar channel. This analysis shows that there are significant differences between the sales channels, which are masked in an aggregated analysis, providing empirical insights into an area where there has been a lack of evidence to this point.

The relevance of the long tail in consumer goods sales, particularly experience goods, has been a subject of extensive discussion over the past two decades. The prevailing idea is that the expanded accessibility of niche or obscure products through online retail, in contrast to the more limited stock of traditional brick-and-mortar stores, leads to greater diversity in consumption patterns. Although each individual product is bought rather little, the vast number of available niche products altogether represent a substantial share of overall sales. While there is no unanimous definition, as discussed below, at least in a common definition according to Brynjolfsson et al. (2011), the 50% least sold products are understood as long tail. While the economic relevance of the long tail is a matter of contention, the latest findings from this debate suggest that the economic relevance of the long tail is probably less than had earlier been assumed. For a review of the relevant literature in this discussion, see the following chapter.

In addition, there are few existing insights in the literature regarding differences in consumption patterns between distribution channels that offer the same products and, in the case of the German book market, even the same prices due to an RPM policy.¹ In this paper, we show that the products purchased via the

¹Since the German “Book Retail Price Maintenance Act” (“Buchpreisbindungsgesetz”) came into force on October 1, 2002, the German book market has been subject to an RPM policy, according to which publishers must fix the retail price of a book for at least the first 18 months after publication. This law also applies to imported books.

respective distribution channels differ significantly.

In contrast to earlier publications in this strand of research, our rich data set allows us to refrain from estimates of sales data. This literature is referred to as the *rank-substitution literature* by Liebowitz and Zentner (2023). In their work, the authors show that this method can result in substantial overestimates of sales, particularly for long tail products and especially for the specifications used in the rank-substitution literature. Instead, we can consider sales data directly. To that end, to the best of our knowledge, this is the first paper to analyze the long tail phenomenon in the context of observed online *and* offline sales data.

The contribution of this paper is twofold. First, we present evidence based on observed sales in the German book market showing that the long tail is of extremely low economic relevance: Depending on the definition applied, its average size varies between 0.93 - 2.45% of overall yearly sales. We then argue against this background that an analysis of the difference between online and offline retail necessitates focusing on the top and mid segment of the sales distribution. That is, products that are considered top sellers as well as products beyond the top sellers but not within the long tail. Using scanner data on the book market in Germany, we show that this mid segment (the “middle tail”²) accounts for a larger share of total sales in online retail.

Second, we utilize our extensive data set on book sales to analyze and provide insights into the distinct composition of products sold through each sales channel. Starting with a high-level distinction between e-commerce and brick-and-mortar retail as it can be found in related literature, our findings reveal a negative relationship between the relative age of books sold through the online and offline retail channels and the relative concentration of the sales in these channels. In particular, our results show that books sold beyond the most selling segment have a lower age in e-commerce than the corresponding books in brick-and-mortar retail, indicating a faster adoption of new titles in e-commerce or, equivalently, a longer survival of successful books in brick-and-mortar retail. Furthermore, we find significant differences in the composition of books titles between sales channels with respect to genres. While belletristic and children’s books play a larger role in brick-and-mortar retail, textbooks make up for a larger share in e-commerce. Our analysis shows that the differences in sales distributions between the sales channels can to a large extent be attributed to differences in the products demanded through the respective channels, i.e., differences in product composition between channels.

In a subsequent analysis we further differentiate the brick-and-mortar channel into chain stores and independent bookstores. From this we show that an aggre-

²To the best of our knowledge, the term “middle tail” was first introduced by Benner and Waldfogel (2023) to describe products between blockbusters and the long tail.

gation of both brick-and-mortar channels masks significant differences between the two store types. Particularly, we present evidence that sales through independent bookstores are less concentrated on top-selling books than those of chain stores. Furthermore, top-selling books in independent bookstores are published more recently and represent a broader set of authors than their counterparts in chain stores, while the opposite is seen in the middle segment.

Overall we can, first, confirm that products in the long tail are of negligible economic relevance. Second, we find evidence for a sales distribution that is less concentrated on most selling titles in online retail and in sales from independent stores in comparison to chain stores. Although these differences show to be rather small, our data reveal a larger share of sales of middle tail products in overall sales through the online sales channel, compared to the chain store sales channel. Finally, we find significant differences in the composition of goods bought across sales channels, in particular by genre. These differences in the book titles consumed can provide an explanation for the limited substitutability of sales channels shown in recent literature.

The remainder of this paper is structured as follows. Section 2 provides a brief overview of literature related to this paper. Section 3 includes a description of our dataset and continues to shed light on the concentration of the sales distribution and size of the long tail. Section 4 presents the results of our comparative analyses of sales channels, pointing out differences in the product compositions between channels. Section 5 concludes.

2 Related Literature

This article is related to the previously mentioned literature on the alleged increase in the relevance of niche products in the context of the digitization of consumer markets. Anderson (2004, 2008) is commonly credited with popularizing the notion that niche products account for a larger share of sales in digital markets than in traditional offline retail, because of a combination of a reduction in production costs and the distribution costs. Around the same time, Brynjolfsson et al. (2003) argue that increased product variety from online retail makes a substantial contribution to consumer surplus. Building on this, Brynjolfsson et al. (2011) explain that online sales have a lower concentration than sales via traditional channels. In a comparison of sales through a “catalog channel” and an “internet channel” of a total of 734 products within one month, the authors come to the conclusion that the long tail (in the sense of the 50% least sold products, Brynjolfsson et al., 2011, p. 1380) is bigger online. In a more recent publication, Brynjolfsson et al. (2022) add to this by presenting evidence that the welfare gains from additional product variety in e-commerce are approximately 40 times

greater than those from price reductions.

However, the economic relevance of the long tail is a matter of contention. To the best of our knowledge, the first to put forward an argument against the economic relevance of the long tail were Elberse (2008) and Elberse and Oberholzer-Gee (2007). Based on evidence from both the home video and music industries they present evidence that the long tail might be of rather little economic relevance. To summarize their conclusions in their own words: “the long tail appears incredibly flat” (Elberse and Oberholzer-Gee, 2007, p. 24). Almost one decade later, Tan et al. (2017) provide supporting evidence that the economic relevance of long tail titles is indeed small. Based on data on movie rentals in the U.S. the authors argue that the demand diversifying effect of additional product variety might affect niche products more strongly than it affects top sellers. According to the authors, the availability of a broader range of products, for example through an online retailer, results in a decrease in the quantities sold of the individual titles. This decrease, however, is pronounced more strongly for niche titles, so that the relative share of the total quantity demanded is shifted in favor of the top sellers. Most recently, Liebowitz and Zentner (2023) show that the method underlying the articles that were central to the acceptance of the long tail effect (in particular Brynjolfsson et al., 2003 and Chevalier and Goolsbee, 2003) produces an overestimation of sales in the long tail. In a general discussion of the appropriateness of double-log linear regression for analyzing the relationship of sales rank and sales quantity, the authors show that this method results in a substantial overestimation of sales quantity for large sales ranks (i.e., in the long tail).

In a related article, Aguiar and Waldfogel (2018) argue that the gains in welfare from additionally available products through the long tail in online retail are even bigger than previous studies suggest. Based on data on the recorded music market, the authors quantify the welfare effects of new music considering the unpredictability of product quality and report effect sizes that are more than 10 times larger than the welfare effects identified in earlier literature.

A brief overview of studies that come to the conclusion that the economic relevance of the long tail is overestimated is already included in the previous section of this paper. A recent paper by Quan and Williams (2018) ties in with this, in which the authors consider the influence of across-market demand heterogeneity on the welfare effects of online retail. Due to brick-and-mortar stores tailoring their product assortment to local demand or local tastes, the relevance of the additional product variety of online retail for welfare is reduced. Overall, the authors find that the welfare effects of additional competitive pressure through online retailing are similar in size to the effects of additional product variety.

Our analysis is also related to the literature on growing relevance of niche

and obscure products in consumption. Employing a model that allows for heterogeneity in demand on a household level, Neiman and Vavra (2023) empirically document a rise in niche consumption. Similar results are presented in Yeo (2022): In an analysis of shopping baskets at household level, they find increasing heterogeneity in household consumption between 2004 and 2016, as well as an increase in the consumption of niche products in the grocery industry. In a comparison of online books sales in 2000 and 2008, Brynjolfsson, Hu, and Smith (2009) present findings consistent with an increase in long tail consumption.

So far, literature examining differences in consumption across sales channels, particularly regarding product preferences or attributes, is scarce. The only study we are aware of in this context analyzes cultural consumption in the realm of museum collections. Navarrete and Borowiecki (2016) compare onsite and online visits to museums and find significant differences in consumer preferences for the type of objects viewed, identifying distinct patterns that reflect the nuances of online and offline consumption. As mentioned earlier, this paper seeks to address this apparent gap in the literature by exploring similar distinctions in consumer behavior across retail channels in the book market.

By highlighting differences in the composition of products purchased online and offline, our work can also inform the literature on the substitutability of sales channels from a consumer perspective. In this context, Liebowitz et al. (2021) provide evidence for the US book market which shows a shift in demand away from bestsellers and towards less successful titles as a result of store closures. In a study of the German book market, Götz et al. (2020) find that brick-and-mortar store closures can explain a decline in overall demand for books, from which they conclude that e-commerce and brick-and-mortar stores are imperfect substitutes. Brynjolfsson, Hu, and Rahman (2009) show that the degree of substitution between e-commerce and brick-and-mortar retail varies based on the market share of the good under consideration. They find that competition is more intense for mainstream goods, while there is limited competition when selling niche products. Similarly, Forman et al. (2009) demonstrate that consumers tend to substitute away from online retail if a physical store exists nearby, attributing this to both the disutility costs of online purchasing and offline transportation costs. Furthermore, Wang and Goldfarb (2017) provide evidence of a coexistence of substitution and complementarity across retail channels. The authors find that in regions where a brand has a strong presence, new store openings reduce online sales, however in regions with weaker brand presence, store openings increase online sales. The authors interpret this as evidence that while “online and offline channels may be substitutes in distribution, they are complements in marketing communications” (Wang and Goldfarb, 2017, p. 706) emphasizing the nuanced relationship between retail formats. Finally, in a study on the US-American book

market, Chevalier and Goolsbee (2003) provide evidence according to which consumers reveal a larger price-elasticity in brick-and-mortar retail in comparison to consumers in e-commerce.

Similar to Liebowitz and Zentner (2023), we have an extensive set of scanner data for the book market we are investigating, which allows us to analyze observed sales. For this reason, we do not need to resort to estimating sales volumes using sales rank based on an observed subset, as is established in the rank-substitution literature (most prominently represented by Brynjolfsson et al. (2003) and Chevalier and Goolsbee (2003)). Based on our data, we can confirm that the relationship between log-rank and log-sales is not linear, as assumed in the rank-substitution literature, but instead concave. We therefore see this work as closely related to the criticism of the rank-substitution method put forward by Liebowitz and Zentner (2023). Similar to the aforementioned study, we also can show that sales in the so-called long tail in consumption are extremely low and were overestimated in earlier literature on this topic.

Finally, this paper is related to the work by Benner and Waldfogel (2023) on the increasing relevance of so-called “middle tail strategies”, according to which producers invest in novel products that are targeted to the preferences of heterogeneous audiences.

3 Concentration of the Sales Distribution

Our analyses are based on scanner data for the German book market including three retail channels: e-commerce, book chain stores (henceforth: chains), and independent bookstores (henceforth: independents). For each retail channel, we observe the top 50,000 book titles that yielded highest revenues for each week from 2011 until and including 2018. We observe each book’s ISBN³, its average price, sold quantity, as well as several characteristics of the respective book. This includes, most notably, its date of release, author(s) and genre. This data was supplemented by historical bestseller lists for each genre, published by the German SPIEGEL magazine. In total, our analyses rely on approximately 55.1 million data points, where each data point consists of weekly sales on an ISBN and sales channel level. This includes data on around 400,000 unique ISBNs per year and a slight over a million unique ISBNs over the entire period. Overall, this data set includes detailed sales data on 87.2% (78.5%) of all books sold through the offline (online) retail channel in Germany in the respective period.⁴ Based

³The *International Standard Book Number* (ISBN) is an internationally used unique identifier for books.

⁴The coverage data was provided by our data provider, the German market research company media control GmbH.

on this data, we aggregate a distribution of sales through the respective retail channel for each year in our observation period, which forms the basis of the analyses in this paper.

Book title consumption exhibits a notable degree of concentration. This concentration is evident when examining the Lorenz curves in Figure 1, which illustrate the extent of disparity in the distribution of book sales in our data. These Lorenz curves reveal two pivotal insights that form the foundation for two central arguments we want to make in this paper. Firstly, all sales channels show a notable degree of concentration. Secondly, Figure 1 seemingly substantiates the commonly asserted notion of a less pronounced concentration within e-commerce compared to brick-and-mortar retail, at least qualitatively. This difference is pronounced more strongly between e-commerce and chain stores. However, it also shows that these differences in sales distributions are rather small. This section continues to discuss the concentration of the distribution of sales, while section 4 offers a detailed analysis of the differences between sales channels, identifying potential contributing factors of these differences in concentrations.

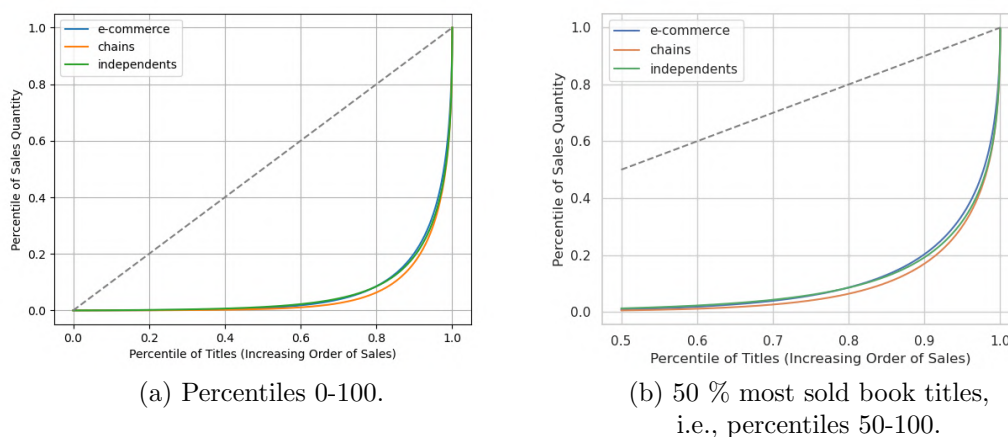


Figure 1: Lorenz curves for yearly book sales, separated by retail channel. Full sales distribution (left) and limited to the 50 % most sold book titles (right).

In order to understand the magnitude of the long tail in our data, we first look at the proportion of total observed sales across all sales channels accounted for by books within each decile of the sales distribution. The results are shown in table 1. One conventional definition of the long tail introduced by Brynjolfsson et al. (2011) includes the bottom 50 % of all products ranked by aggregated sales. Table 1 shows that the aggregate share of all titles within the long tail according to this definition constitutes, on average, approximately 1.63 % of the overall sold quantity of books in each year.⁵ Following an alternative, more narrow definition

⁵An analysis based on revenues instead of sold quantities yields an unchanged result and is included in the appendix.

of the long tail discussed in related literature, which delimits the long tail as the bottom 40 % of sold products, this average share diminishes to below 1 % (more precisely, 0.93 %). This stark reduction underscores the sensitivity of long tail metrics to definitional nuances. More importantly, both Figures emphasize the limited economic relevance of the long tail in book consumption, signaling that the vast majority of sales are concentrated within the top segments of the distribution.

These observed sales shares are considerably lower than those reported in earlier publications. In particular, these shares are in contrast with the results of the rank substitution literature, for example Brynjolfsson et al. (2011), who report shares of 12.7 % and 15.2 % for the bottom 50 % in the data they examined for catalog sales and online sales, respectively. Our observations are more in line with those of Liebowitz et al. (2021). In their analysis of the US book market, the authors show cumulative sales of around 2 % to 4 % for the 75-90 % least sold books. This implies that the US book market is concentrated even stronger than the German market where, according to our data, the lower 80 % of all books sold comprise 10,2 % of all sales (see Table 1 below).

In the discussion about the role of the long tail, there is an alternative definition in absolute terms. According to this, all products that exceed the 100,000th rank are considered part of the long tail. This measure was first proposed by Brynjolfsson et al. (2003) who motivate it by pointing out that this would correspond with the approximate stock size of a large book store (“superstore”) in the U.S. Applied to our data on the German book market, the long tail according to this absolute definition would account for a slightly larger share of total sales compared to the previously discussed relative definitions (e.g., 50 % least sold books). Our data shows that all titles beyond the 100,000 most sold books are responsible for an average of 2.45 % (median: 2.73 %) in overall sales per year.

In previous publications, it is argued that papers finding an importance or growth of the long tail typically use absolute definitions of the long tail while papers coming to the opposite conclusion would typically use a relative definition (Brynjolfsson et al., 2010). On the basis of the sales shares presented, which we identify in our data on the German book market, the overall picture is that the long tail is extremely small in both absolute and relative terms.

A similar conclusion can be drawn when considering the Gini coefficients. Brynjolfsson et al. (2011) report Gini coefficients of 0.49 and 0.53 for online and catalog retail channels, respectively. The Gini coefficients identified by Hinz et al. (2011) for the online video-on-demand market are slightly larger, ranging from 55.3 to 73.2 (mean value: 63.7). Our data unveil substantially larger Gini coefficients of 0.87 and 0.89 for the online and both offline channels combined while the Gini coefficients for chains and independents are 0.89 and 0.88, respectively.

Decile	Share in Sales [in %]	Decile	Share in Sales [in %]
1	77.78 (3.84)	6	0.70 (0.13)
2	12.07 (1.74)	7	0.42 (0.07)
3	4.92 (1.11)	8	0.26 (0.04)
4	2.36 (0.60)	9	0.17 (0.04)
5	1.24 (0.29)	10	0.08 (0.02)

Table 1: Share of sales of the respective decile of books in overall sales of books. The displayed values represent the mean share (standard deviation) of the respective decile over the total observation period (i.e., 2011-2018). Yearly data.

These observations are consistent with those of Tan et al. (2017), who state Gini coefficients between 0.82 and 0.86 in their study of physical video rentals.

We find our observations to be generally in line with those of the recent literature on the long tail in digitized markets. Liebowitz et al. (2021) and Liebowitz and Zentner (2023) in particular argue that earlier studies underestimate the concentration of sales in the book market and overestimate the long tail. Similar to our work, both papers are based on scanner data, although we analyze data for the German and Liebowitz et al. (2021) and Liebowitz and Zentner (2023) analyze data on the U.S. book market.

As evident from the Lorenz curves presented earlier in Figure 1, differences in concentration between sales channels appear to be relatively subtle. Figure 2 shows that these differences in the concentration of sales channels decrease in size along the sales distribution. While for the first decile the difference in median share in sales between the most concentrated channel, i.e., chains and the least concentrated channel, i.e., e-commerce, is 8.71 percentage points, this difference diminishes to below one percentage point for deciles 5 and above. The differences between retail channels in sales shares in the lower half (i.e., the long tail) are negligible. At the same time, Figure 2 shows that the sign of these differences changes between the top 10 % (i.e., most sold books) and the 11-50 percentile (i.e., middle segment). As a result, brick-and-mortar retail shows a stronger concentration on the 10 % most sold titles. This is pronounced stronger for chains as it is for independents. In contrast, online retail reveals higher sales shares in the middle segment in comparison to both brick-and-mortar retail channels. These differences between sales channels motivate our more detailed discussion of the dissimilarities between the products in each sales channel in the following section.

In a comparison of sales concentrations over time, our data provide no indi-

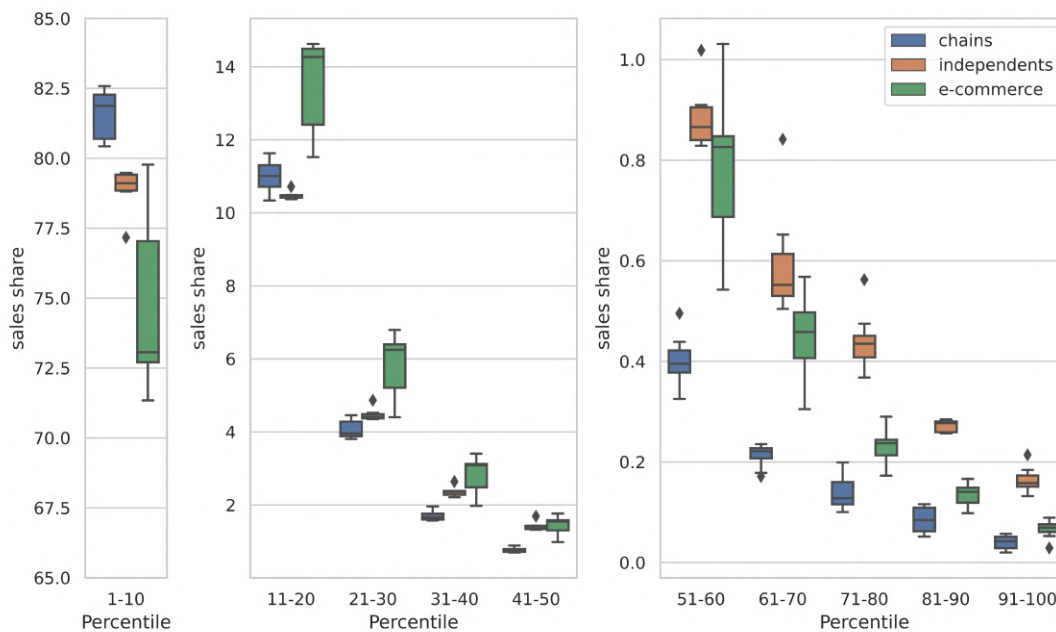


Figure 2: Share of quantity sold within specified percentiles of the sales distribution in total quantity sold, separated by sales channels.

cation of a shift. In particular, there is no evidence of an increase in the sales shares of the last decile (i.e. the long tail), as suggested by the literature cited above. A plot of the sales shares of the books within the respective deciles for the period from 2011 to 2018 is included in the appendix (Figure A.1).

In conclusion, the presented analysis of the long tail in consumption of books in Germany cannot provide compelling evidence supporting overall economic relevance of the long tail. Moreover, it does not offer indications aligning with a notably stronger economic relevance of the long tail in online retail compared to offline retail. Our following section will delve into this matter, demonstrating that differences in concentration between retail channels are primarily driven by disparities in the best selling segment and the middle segment, i.e., titles beyond most selling books and before the long tail.

4 Compositional Differences Between Retail Channels

After introducing the metrics on which our analysis is based, subsection 4.1 highlights the differences between e-commerce and brick-and-mortar retail as a whole. Subsection 4.2 then continues to show that a distinction between independent bookstores and chain stores reveals striking differences between these two channels that are lost in the aggregation of the brick-and-mortar channel.

Our data reveals a notable difference in size between retail channels. A dis-

distinction between e-commerce and the entire brick-and-mortar retail shows that e-commerce only accounts for approximately 22 % of the total trade volume. A more detailed breakdown of brick-and-mortar retail into chains and independents furthermore shows that of the sum total of 78 % that brick-and-mortar retail accounts for in total, independents make up for approximately 50 % and chains circa 28 %. Given these considerable differences in size between the sales channel, we find it advantageous to assess differences between sales channels across the sales distribution by examining the share of sales within percentiles of the sales distribution relative to the overall sales for each channel. This metric is further normalized to be mean zero to adjust for differences in levels between (intervals of) percentiles in sales ranks. Equation 1 captures this transformation.

$$\sigma_{c,s} = \frac{\sum_{i \in s} sales_{i,c}}{\sum_i sales_c} - \frac{\sum_{i \in s} sales_i}{\sum_i sales} \quad (1)$$

In Equation (1), s denotes the segment of the sales distribution. Particularly, s represents (an interval of) percentiles of the sales ranks any observed book i falls into. Index c represent the sales channel. From this definition, it follows that $\sigma_{c,s} \in (-1, 1)$. While this metric is calculated for each period t in our data, the time subscript is dropped for the sake of simplicity. Specifically, an exemplary data point in the first decile for the brick-and-mortar category with a value of around $\sigma_{B\&M,1} = 0.04$ implies that in brick-and-mortar retail, the share of the 10 % most selling books in a year was four percentage points (p.p.) higher than the share of the 10 % most selling books across all channels. In this case, therefore, there is a higher concentration of sales on the top 10 % most selling books in brick-and-mortar retail relative to total sales through all retail channels.

4.1 Comparison of e-Commerce and Overall Brick-and-Mortar Retail

Having established the metrics used for the analyses, a comparative analysis of e-commerce and brick-and-mortar retail as a whole, i.e. both independent stores and chain stores, is carried out. Figure 3 shows these differences in sales shares for segments of relative sales ranks defined by deciles of the sales distribution and for a distinction between e-commerce and brick-and-mortar retail as a whole. It should be noted that this analysis of just two sales channels simplifies the metrics presented above. In this case, the difference between the two sales channels can be read directly, without the need for a comparison with the mean value across all channels.

Notably, the differences in sales shares $\sigma_{c,s}$ of deciles undergo discernible shifts along sales ranks. The 10 % most selling book titles constitute 0.6-4.8 p.p. (mean:

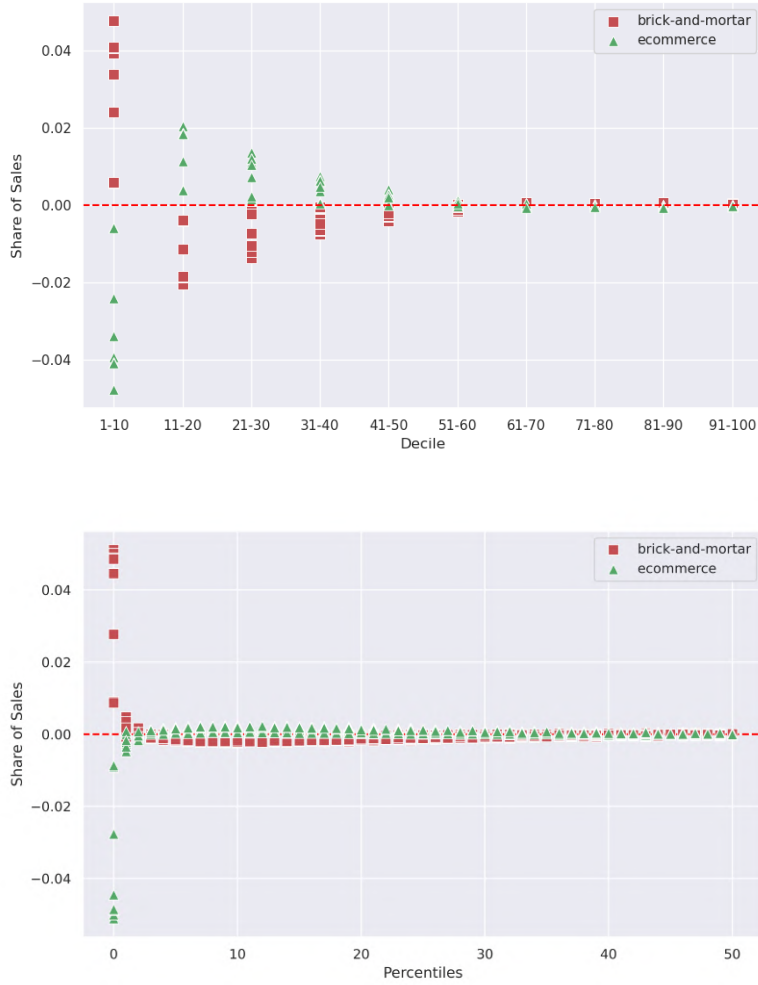


Figure 3: Sales shares within segments of relative sales ranks across sales channels for two sales channels: e-commerce and brick-and-mortar stores. Calculated separately for deciles over the whole distribution of sales (upper plot) and percentiles over the 50 % most selling titles (lower plot). Based on yearly sales, i.e., each dot represents $\sigma_{c,s}$ for yearly sales data.

2.98 p.p.) more of total sales in the brick-and-mortar sales channel than total sales through both channels within this segment. In this distinction between two sales channels only, this is equivalent to a 1.2-9.6 p.p. difference between both channels. This differences shows to be largely driven by the top 1 % most sold book titles (lower plot in Figure 3). Conversely, this trend inversely applies to subsequent deciles from the 11-20 % to the 41-50 % most selling titles (i.e., the middle segment). Here, the share in overall sales is greater in the e-commerce channel as exemplified by the second decile with a difference of 0.9-4.1 p.p. (mean: -2.9 p.p.) between both channels under consideration. This difference is less pronounced for books sold less often, i.e., with larger sales rank. Finally, for the 50 % least sold titles (percentiles 51-100) no discernible differences between sales channels are evident as the mean difference for the 51-60 % decile is already

negligible at 0.1 p.p.

The key insights from this are twofold. First, the near zero difference between sales shares between sales channels for the 50 % least sold books, conventionally regarded as the long tail in consumption, challenges the prevalent belief of a more pronounced long tail effect in e-commerce. Our data does not affirm the presumed prominence of the long tail in the online retail. Second, we do observe larger shares in sales through the online retail channel in the second through fourth deciles. This observation resonates with the widely acknowledged more general notion of a less concentrated sales distribution in online retail. These findings suggest that, contrary to earlier literature, the focus of analysis comparing traditional brick-and-mortar sales with online sales should not primarily be on the long tail. Instead, attention should be directed towards the middle segment, which comprises book titles beyond bestsellers but still sold more often than the lower 50 % of the sales distribution.

Our comprehensive dataset enables a more detailed examination of differences between sales channels. This allows us to analyze variations in book characteristics across channels and, thereby, differences in product composition between sales channels. The differences in product characteristics are calculated in an analogous way to the calculation of sales shares described above: separated for each sales channel and relative to total sales through all channels as well as normalized to a mean of zero.

Specifically we consider the following characteristics: First, the mean number of authors is considered as a measurement for the degree of diversity in sold books. This is calculated by putting the total number of unique authors in the segment of the sales distribution and sales channel combination under consideration in relation to the total number of books in the same segment and sales channel combination. Second, the mean age of book titles within each segment-channel combination is calculated as the mean number of days since release of a book. This measure can be understood as an indicator of the timeliness of the titles consumed or the speed of adoption of new titles in the respective channel. In particular, this can provide an indication of the extent to which new titles are “discovered” by consumers in a distribution channel or whether less recent, already established titles are primarily consumed.

Similarly, third, the share in books that are classified under each of the considered genres is calculated for each segment-channel combination. Insofar as books of different genres can be understood as imperfect substitutes, differences in the composition of the genres may indicate that the distribution channels under consideration serve different parts of book demand. This metric is calculated for genres belletristic, children’s books, travel & guide books, textbooks, and non-fiction books. Finally, the share of (previously) bestsellers is calculated for each

segment-channel combination where a book is considered a bestseller, if it made an appearance on a published bestseller list in any period previous to the time of observation. In this, genre specific bestseller lists are taken into consideration in order to control for differences in market sizes between genres.

These differences then form the basis for a linear regression of a third-degree polynomial. Figure 4 shows these differences in product characteristics between the sales channels and across the sales distribution, as well as the fit of a third-degree polynomial with a 99 % confidence interval portrayed in the shaded areas. In this, the x-axis represents rank percentiles, while the y-axes portray the differences between sales channels concerning the respective characteristics of the sold books.

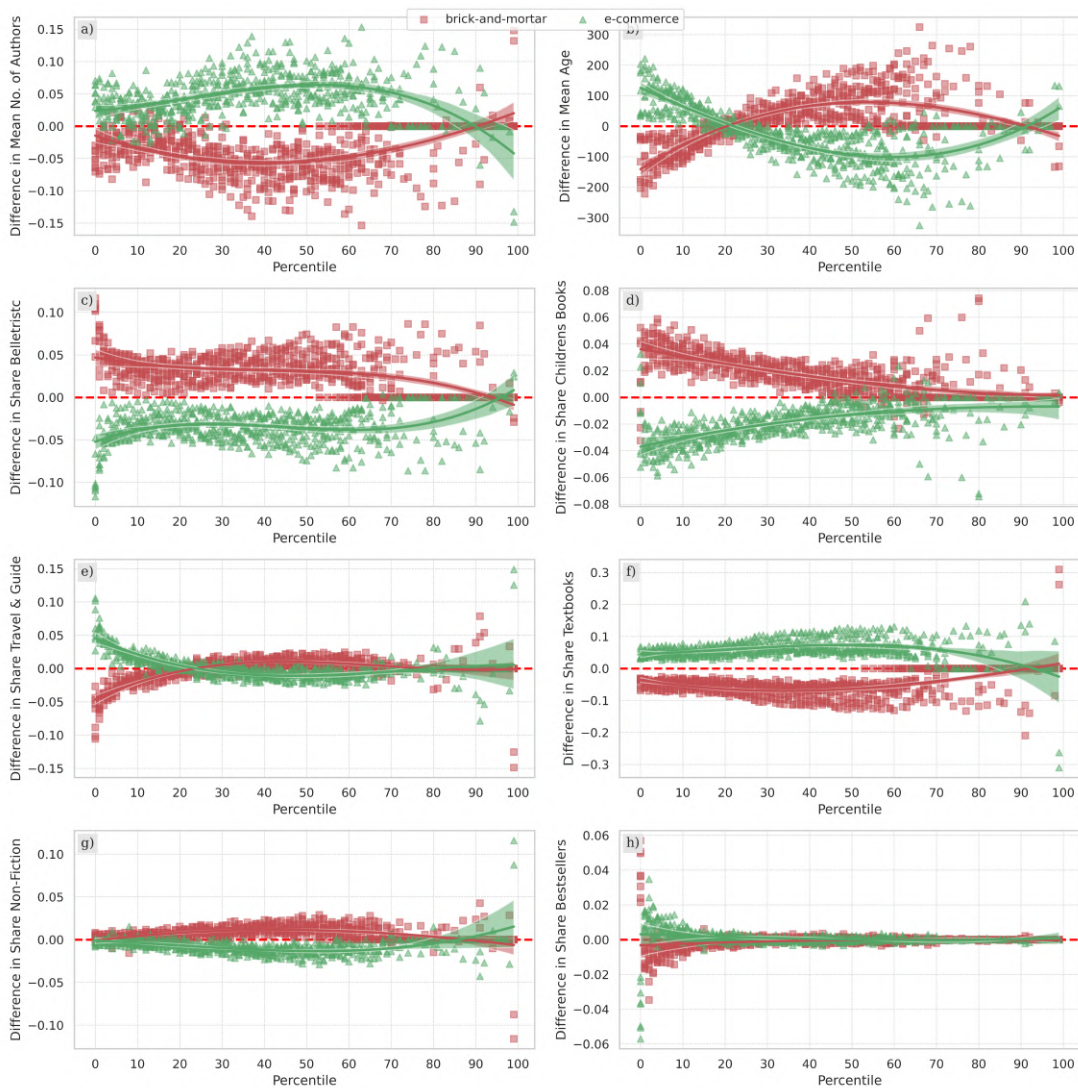


Figure 4: Differences in books' characteristics between sales channels and across the sales distribution. The solid lines shows the fit of a third degree polynomial linear regression model with 99% confidence intervals represented by the shaded areas.

First, panel a) of Figure 4 reveals that the online retail channel features a

significantly larger number of authors by between 2.5-6.5 p.p., considering the upper 80 % of the sales distribution and in comparison to the mean over both sales channels. Notice, as this analysis is limited to two distinct retail channels only, this implies a difference between both channels of 5-13 p.p. While it could be argued that this difference suggests a more diversified consumption through the e-commerce channel, this difference might be primarily driven by variations in the genre composition of books across sales channels, a point to be elaborated in the following. It can be shown that genres more prominent in the top-selling percentiles in brick-and-mortar stores tend to be co-written by fewer authors as discussed in detail below in the context of the analysis of data on genre level.

Second, panel b) of Figure 4 shows that there is notable differences in the mean age of books between sales channels. Top selling books in the online sales channel reveal to be markedly older, approximately by 200 days, compared to their counterparts in brick-and-mortar stores. This indicates that the most sold titles are more topical in brick-and-mortar in comparison to e-commerce. Conversely, a similar age discrepancy in the opposite direction is observed for books in the middle segment of the sales distribution.

Third, the second to fourth rows of panels in Figure 4 (particularly, panels c)-g)) show the differences between the sales channels in the composition of the genres. It can be shown that both belletristic (5-10 p.p.) and children's books (2.5-7.5 p.p.) constitute significantly larger shares in brick-and-mortar sales in comparison to the e-commerce retail channel. Conversely, the opposite situation is evident for books categorized under the textbook genre, where e-commerce sales command a significantly larger share in comparison to sales through brick-and-mortar stores (approx. 10 p.p.). These findings suggest genre-specific consumer preferences on the choice of retail channel that contribute to the observed differences in sales distribution across channels.

Finally, the bottom right panel h) of Figure 4 compares the number share of books that have appeared on a bestseller list in any period previous to the time of observation. Distinct differences can only be seen here for the 1 % most-selling titles. Among these, former bestsellers make up a noticeably larger proportion (between 1-7 p.p.) in brick-and-mortar retail than in e-commerce. An opposite effect can be suspected for the remainder of the percentiles in the first decile, however, these differences are not significantly different from zero.

The overall picture that emerges is that, first, a broader group of authors is represented in online retail in the relevant segments of terms of sales volumes. Second, most-selling books in online retail are younger, and third, more textbooks, travel guides and other guide books are bought in this segment in online retail, while more belletristic and more children's' books are bought in brick-and-mortar retail. Finally, it can be seen that although books from the

bestseller list make up a larger proportion of the top 1 % most-selling books in brick-and-mortar retail, their share is larger in online retail in subsequent percentiles of the sales distribution.

In light of the previously discussed hypothesis that variations in book characteristics between sales channels are driven by underlying differences in sales composition, particularly evident in differences in genre distributions, a more granular analysis at the genre level is presented in the following. Figure 5 shows the results of this analysis for those genre-subsets of our data exhibiting the most significant differences in their share in overall sales, namely belletristic, children’s books, and textbooks.⁶

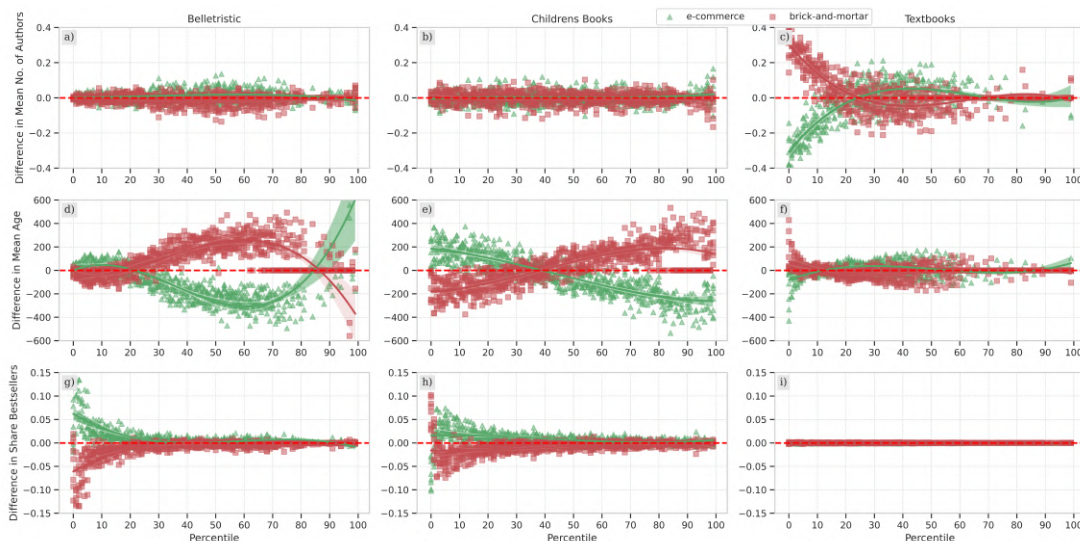


Figure 5: Differences in books’ attributes for different genre subsets. The solid line shows the fit of a third degree polynomial linear regression model with 99 % confidence intervals represented by the shaded areas.

Considering Figure 5, it becomes apparent that differences between sales channels are generally of a smaller magnitude in this genre specific analysis. Nevertheless, notable exceptions show:

First, within the belletristic genre (first column in Figure 5, differences in the number of authors between channels are no longer significant (panel a). The same result shows for the difference in age among the most selling books in this genre (panel d). However, the difference in age throughout the middle segments discussed in the previous analysis of all genres combined remains: Middle tail books sold through e-commerce reveal to be younger by up to one year. In contrast, the differences in the number of previous bestsellers among the most selling titles (panel g) show the opposite pattern from the previous analysis on all genres. In this subset of books in belletristic, the share of previous bestsellers among the most selling titles is significantly larger in e-commerce.

⁶Results for all other genres are provided in the appendix A.2.

Second, considering children’s books (second column in Figure 5, again, differences in the number of authors are no longer significant (panel b). Differences in mean age (panel e) and in the share of previous bestsellers (panel h) reveal generally the same structure as discussed in the previous analysis of all genres combined. Particularly most selling books in brick-and-mortar are published more recently, and furthermore the top 1 % most selling titles contain more previous bestsellers in the brick-and-mortar channel while the subsequent percentiles reveal the opposite.

Third, significant differences in the number of authors are observed for textbooks (third column in Figure 5, panels c)), particularly notable in the most-selling percentiles. Notably, most-selling textbooks in brick-and-mortar retail are authored by a larger group compared to those in e-commerce, a property masked in in previously discussed analysis for all genres combined. For both other measures, no significant differences can be observed.

Overall, these findings support the hypothesis posited earlier, suggesting that differences in sales distribution and product characteristics such as author count, age since release, or the prevalence of established bestsellers are substantially driven by variations in genre composition within each retail channel. Notice that in Figure 5 for both belletristic and children’s books, established bestsellers from previous years command a larger share in top-selling titles within online retail.

Generally, differences in the composition of purchased book titles between sales channels are consistent with the previously described paper by Quan and Williams (2018) who in their analysis of a clothing market also find evidence of adjustments to the assortment of brick-and-mortar stores to across-market demand heterogeneity. Furthermore, this result is consistent with the findings of Navarrete and Borowiecki (2016) described above, who find evidence of differences in preferences between distribution channels.

4.2 Comparison of e-Commerce, Chain Stores and Independent Stores

Following the discussion in the previous section on the differences between e-commerce and overall brick-and-mortar retail, this section further distinguishes the brick-and-mortar sales channel into chain stores and independent bookstores. This differentiation aims to provide a more nuanced understanding of the sales distribution within the traditional offline retail sector.

Figure 6 depicts the sales shares within deciles of the sales distribution. The presentation method mirrors that used in the previous subsection, particularly in Figure 3. Specifically, it illustrates the share of sales within each channel attributable to the respective percentile intervals, contrasted with the share of sales

across all channels for the same percentile intervals. As observed in Figure 3 in the previous section, this Figure also indicates that online retail is less concentrated on the most selling books (i.e., the first decile) compared to brick-and-mortar retail. Within this detailed analysis of brick-and-mortar retail, it additionally becomes evident that chain stores exhibit a markedly higher concentration on the highest-selling books compared to independent book stores. Overall, it appears that differences in sales between online and offline retail are largely driven by differences in sales between chains and e-commerce. Particularly, the stronger concentration on most selling titles visible in brick-and-mortar retail is mostly seen in chains.

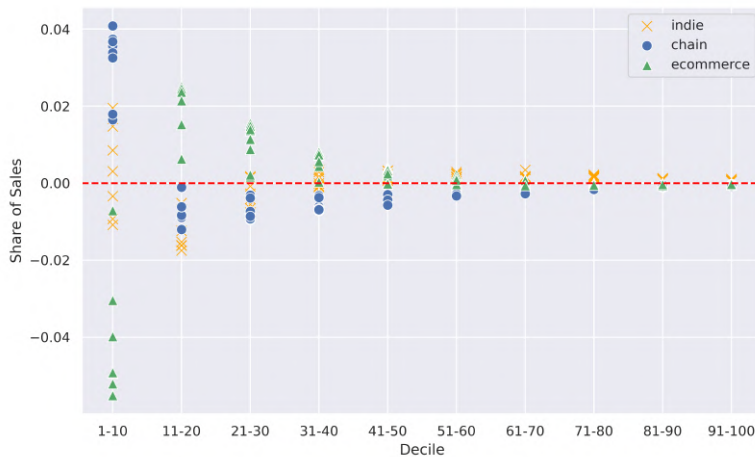


Figure 6: Sales shares within segments of relative sales ranks across sales channels for three sales channels: e-commerce, chain stores and independent stores. Based on yearly sales, i.e., each dot represents $\sigma_{c,s}$ for yearly sales data as defined in Equation 1.

Similar to our discussion in the previous subsection, we will now examine differences in various characteristics of books across the distribution channels. The underlying data analysis is analogous to that explained in connection with Figure 4 in the previous subsection. Figure 7 presents the results of this evaluation. It should be noted here that, as shown in Section 3, only the upper deciles (i.e., the most-selling books and middle segments) are of economic relevance. For this reason, the focus of the following interpretation of the analysis is on these segments.

While it was previously shown that the average number of authors is higher in e-commerce across all relevant segments compared to brick-and-mortar retail, this more granular analysis reveals a more nuanced picture. Specifically, there is no significant difference in the number of authors for highest selling books in e-commerce and independent stores (see panel a) in Figure 7). However, the relative number of authors represented in independent stores decreases rapidly

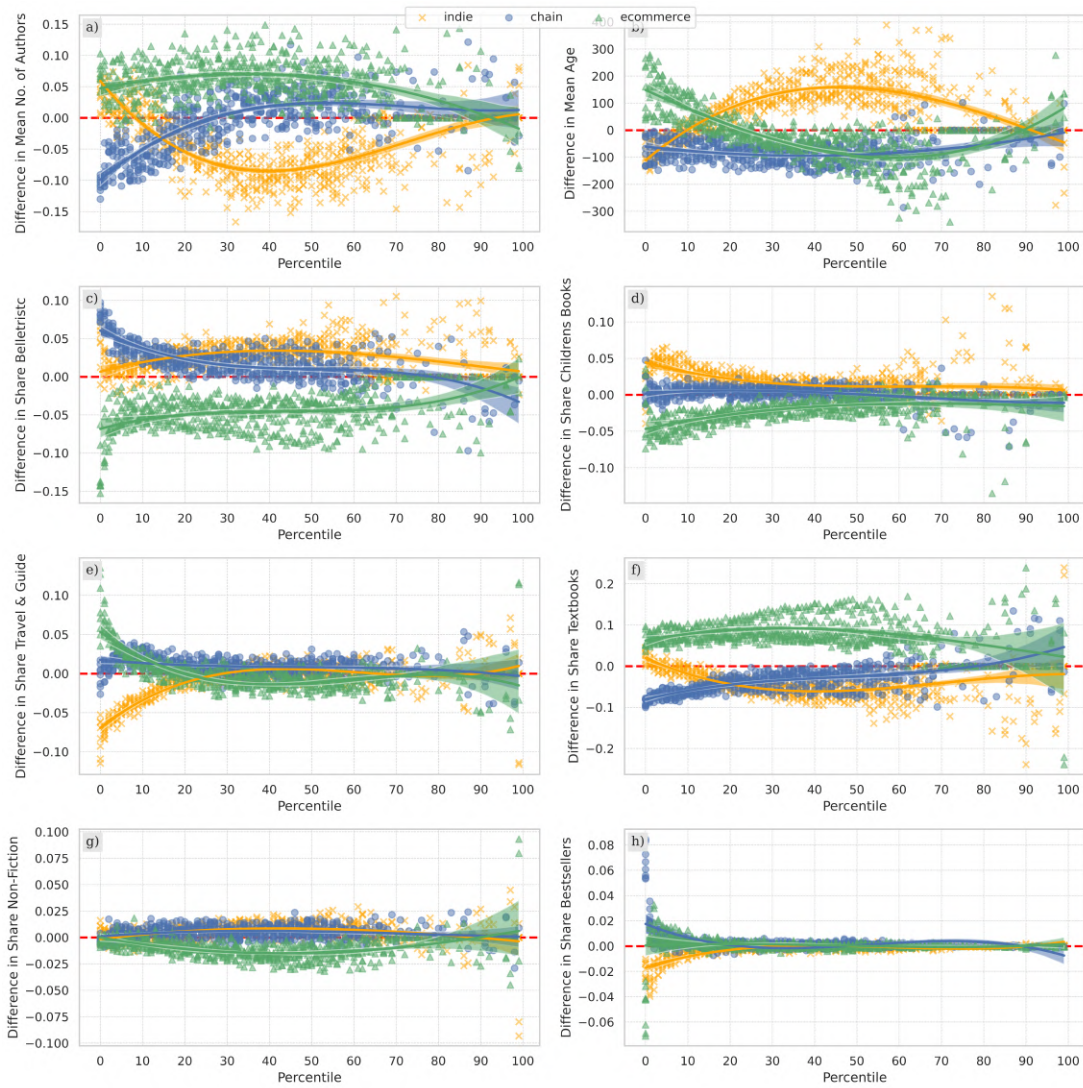


Figure 7: Differences in books’ characteristics between sales channels and across the sales distribution. The solid lines show the fit of a third degree polynomial linear regression model with 99% confidence intervals represented by the shaded areas.

beyond the highest selling decile and reveals a negative sign throughout the middle segments (decrease from approx. 5 p.p. to -7 p.p.). Overall, the breadth of represented authors for the highest-selling books is significantly lower in chain stores compared to both e-commerce and independent book stores at a difference of up to 15 p.p. As argued earlier, interpreting this difference requires considering genre composition differences. Still it seems clear that both e-commerce and independent stores distribute a more diverse group of authors within the most selling books segment. This segment is particularly relevant as it accounts for roughly three-quarters of all sales (see table 1). Additionally, panel a) indicates that e-commerce represents more diverse authors across the entire sales distribution, whereas independent stores feature significantly fewer authors in the middle segment.

Regarding the average age of books (panel b), notable differences between the two brick-and-mortar channels are observed. The pattern identified in the previous section, where the most-selling books in brick-and-mortar retail are significantly younger than those in e-commerce, is also apparent at the level of chain stores and independent stores. However, throughout the remaining sales distribution, significant differences in the age of books between independent stores and chain stores are evident: While books sold in the middle segments in independent stores are significantly older than average by circa half a year, those sold in chain stores are younger than average across all segments by up to 100 days.

The analysis of genre composition furthermore reveals significant differences within the brick-and-mortar channels. The above-average presence (~ 6 p.p.) of belletristic in the most selling segment is driven by chains (panel c). In contrast, panel d) shows that children’s books hold a higher share in independents (~ 5 p.p.), whereas their presence in chain stores is not significantly different from the average over all channels. Lastly, panel f) highlights the differences in concentration within the textbook genre: while textbooks have a larger share in the most selling segment of independent stores compared to chain stores (~ 10 p.p.), this ratio reverses in the middle segments.

Panel h) shows the proportion of former bestsellers in total sales. Differences between the channels are observed only for the most selling books, similar to the comparison in the previous subsection. Here, it is evident that bestsellers constitute a significantly above-average share in chain stores (~ 2 p.p.). In independent stores, the opposite is true: the share is significantly below average to a similar extent.

Overall, this analysis highlights three main points: Firstly, a combined consideration of the entire brick-and-mortar retail sector is inadequate for analyzing the differences between distribution channels. Instead, it is clear that the differences between chain stores and independent stores are significant and relatively large. Secondly, significant and pronounced differences between the distribution channels are evident for both the most selling books and the middle segments, while the least sold books (i.e., the long tail of the sales distribution) exhibit relatively small and mostly insignificant differences.

Thirdly, it can be seen that the differences in the concentration of distribution channels shown in section 4 are linked to differences in the composition of the books sold. Specifically, within the most selling books, (i) chain stores sell an above-average number of previous bestsellers and titles of genre belletristic while selling below-average numbers of textbooks, which is accompanied by a lower author diversity. (ii) Independent book stores, on the other hand, sell a below-average number of previous bestsellers and show a concentration on children’s books within the most selling segment, with guide & travel books being

underrepresented. In the middle segment, there is an above-average share of belletristic and a below-average share of textbooks. (iii) E-commerce qualitatively retains the same characteristics described in the previous section 4: across the entire sales distribution, it represents a significantly higher number of authors, with belletristic and children’s books being underrepresented and textbooks being overrepresented.

Analogous to our analysis in section 4.1, our data allows us to analyze the differences in the characteristics of sold books while controlling for genre effects. Figure 8, similar to the approach in Figure 5, shows selected characteristics for those genres where the distribution channels differ the most.⁷ Figure 8 shows that differences in the characteristics of sold books between distribution channels are less pronounced and, in some cases, no longer significant, once we control for genre. This is particularly pronounced for the number of authors (first row in Figure 8). Only for textbooks (panel c) does the number of authors online remain significantly lower than the average across all channels for the most selling titles. Additionally, it is evident that textbooks in the independent store channel represent an above average variety of authors. This can be interpreted as an indication of greater product diversity in the textbook category within the independent bookstore channel. Overall, this analysis confirms that differences in author diversity between distribution channels are largely influenced by differences in genre composition.

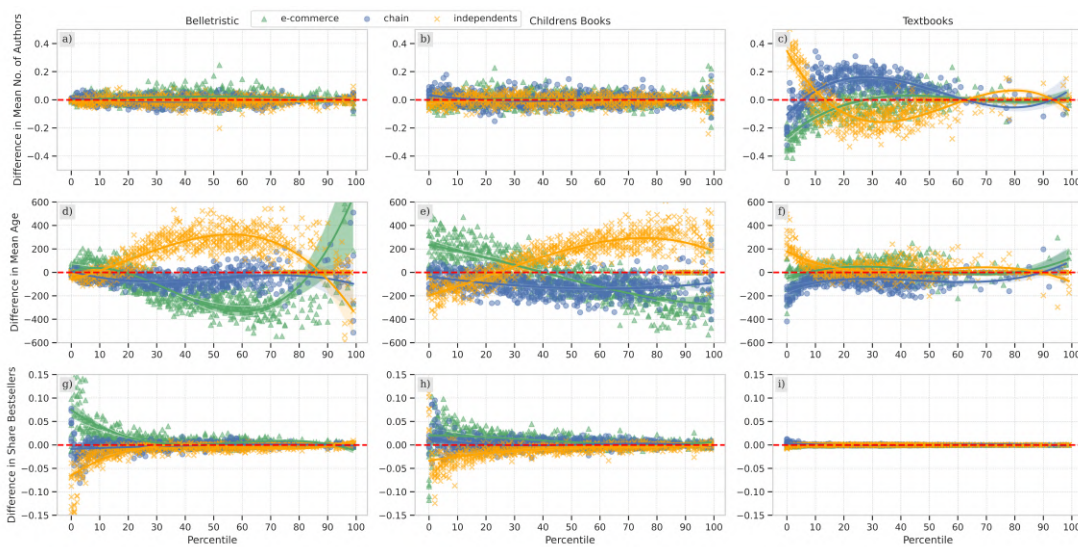


Figure 8: Differences in books’ attributes for different genre-subsets. The solid line shows the fit of a 3rd degree polynomial linear regression mode with 99% confidence intervals represented by the shaded areas.

Regarding the age structure of sold books (second row in Figure 8), it becomes clear that differences between the channels mainly arise from e-commerce and in-

⁷Results for all other genres are provided in the appendix A.3.

dependent book stores: while books sold through independent stores, particularly in the middle segment, have a significantly higher average age, the opposite is true for books sold through e-commerce.

A similar pattern is seen in the third row of Figure 8, which shows the share of previous bestsellers in total sales. Here, it is evident that the difference between online and overall brick-and-mortar retail is again driven by differences between independent stores and e-commerce: former bestsellers constitute a significantly higher share in e-commerce compared to independent book stores.

While these insights are generally consistent with the observation that differences between distribution channels are largely driven by differences in the genres of sold books, some differences between channels within genre still remain. In this genre-specific analysis, it is shown that in the most selling books category significantly more established bestsellers are purchased in e-commerce. This appears to be generally in line with findings of related work on the book market, according to which sales volumes in e-commerce increase after a book appears on a bestseller list, whereas the opposite correlation can be seen in brick-and-mortar retail (Goetz et al., 2020). Additionally, independent book stores tend to sell older books on average. Moreover, it can be argued that sales of independent bookstores are stronger concentrated on books beyond already established bestsellers.

5 Conclusions

With the analyses presented in this paper we aim to contribute to the ongoing discussion on the effects of digitization on consumer goods markets. The main contributions of this work are threefold: First, we provide evidence confirming little economic relevance of the long tail. Furthermore, we cannot confirm significant differences between sales channels regarding the relative relevance of the long tail. Instead, relevant differences between sales channels can be identified for most selling titles as well as the middle segments (i.e., the “middle tail”) of the sales distribution. An analysis of these segments of the sales distribution reveals that the long discussed increased consumption in titles beyond the bestsellers through the online retail channel indeed prevails in this middle tail.

Second, we provide a nuanced picture on the differences in the compositions of sales between sales channels. Drawing from our rich set of scanner data, we are able to establish that these differences in the sales distribution are related to differences in the characteristics of the books sold through the respective channel and in the respective segment of the sales distribution. In particular, we show that books sold in the middle segment, which is stronger pronounced in online retail, are older in the brick-and-mortar channel. This is in contrast with the idea

that the online retail channel might merely represent a backlog of book titles no longer available in brick-and-mortar bookstores and suggests the opposite: Segments of the sales distribution that are stronger in the online retail (i.e., the middle segments) show to contain younger books in the online retail channel than in brick-and-mortar retail. Overall, our analyses indicate that these differences between the distribution channels are to a large extent due to differences in genre composition.

Third, a more disaggregated analysis is presented in which we distinguish not only between e-commerce and brick-and-mortar retail, but furthermore distinguish between chain stores and independent book stores within the brick-and-mortar channel. This analysis reveals stark differences between the composition of sales between chain stores and independent book stores which otherwise are masked by an aggregation into brick-and-mortar retail as a whole. The differences thus discernible indicate that book consumption through independent book stores is more diverse than consumption through chain stores. In particular, it can be seen that the share of established bestsellers in overall sales in independent stores is lower than in the other sales channels.

This paper thus fills a gap in the existing literature by highlighting differences in the composition of products sold through different channels. Our results suggest differences in consumer preferences between sales channels. Particularly, we show significant differences in the composition of genres between the sales channels, which could indicate that different types of consumers are served by the respective sales channels, implying that e-commerce and brick-and-mortar in general as well as chain stores and independent stores are imperfect substitutes. We argue that this finding is consistent with and supported by recent advances in the literature finding limited substitutability between sales channels and channel specific preferences.

References

- Aguiar, L., & Waldfogel, J. (2018). Quality predictability and the welfare benefits from new products: Evidence from the digitization of recorded music. *Journal of Political Economy*, 126(2), 492–524.
- Anderson, C. (2004). The long tail. *The Wired Magazine*, Outubro (October), 170–177.
- Anderson, C. (2008). *The long tail: Why the future of business is selling less of more*. Hyperion. http://www.worldcat.org/search?qt=worldcat_org_all&q=9781401309664.

- Benner, M. J., & Waldfogel, J. (2023). Changing the channel: Digitization and the rise of “middle tail” strategies. *Strategic Management Journal*, *44*(1), 264–287.
- Brynjolfsson, E., Chen, L., & Gao, X. (2022). *Gains from product variety: Evidence from a large digital platform* (Working Paper No. w30802). National Bureau of Economic Research.
- Brynjolfsson, E., Hu, Y., & Rahman, M. S. (2009). Battle of the retail channels: How product selection and geography drive cross-channel competition. *Management Science*, *55*(11), 1755–1765.
- Brynjolfsson, E., Hu, Y., & Simester, D. (2011). Goodbye pareto principle, hello long tail: The effect of search costs on the concentration of product sales. *Management science*, *57*(8), 1373–1386.
- Brynjolfsson, E., Hu, Y., & Smith, M. D. (2003). Consumer surplus in the digital economy: Estimating the value of increased product variety at online booksellers. *Management science*, *49*(11), 1580–1596.
- Brynjolfsson, E., Hu, Y., & Smith, M. D. (2010). Research commentary—long tails vs. superstars: The effect of information technology on product variety and sales concentration patterns. *Information Systems Research*, *21*(4), 736–747.
- Brynjolfsson, E., Hu, Y. J., & Smith, M. D. (2009). A longer tail?: Estimating the shape of amazon’s sales distribution curve in 2008. *Workshop on Information Systems and Economics (WISE)*, 1–6.
- Chevalier, J., & Goolsbee, A. (2003). Measuring prices and price competition online: Amazon. com and barnesandnoble. com. *Quantitative marketing and Economics*, *1*, 203–222.
- Elberse, A. (2008). Should you invest in the long tail? *Harvard business review*, *86*(7/8), 88.
- Elberse, A., & Oberholzer-Gee, F. (2007). Superstars and underdogs: An examination of the long-tail phenomenon in video sales. *MSI Reports: Working Paper Series*, *4*, 49–72.
- Forman, C., Ghose, A., & Goldfarb, A. (2009). Competition between local and electronic markets: How the benefit of buying online depends on where you live. *Management science*, *55*(1), 47–57.
- Goetz, G., Herold, D., Klotz, P.-A., & Schäfer, J. T. (2020). Innovation, bestsellers and digitization—where to find the needle in the haystack? *MAGKS Joint Discussion Paper Series in Economics*, (No. 12-2020).
- Götz, G., Herold, D., Klotz, P.-A., & Schäfer, J. T. (2020). The substitutability between brick-and-mortar stores and e-commerce—the case of books. *MAGKS Joint Discussion Paper Series in Economics*, (No. 11-2020).

- Hinz, O., Eckert, J., & Skiera, B. (2011). Drivers of the long tail phenomenon: An empirical analysis. *Journal of management information systems*, 27(4), 43–70.
- Liebowitz, S. J., Ward, M. R., & Zentner, A. (2021). Only a “longish” tail. *Available at SSRN 3770454*.
- Liebowitz, S. J., & Zentner, A. (2023). The challenges of using ranks to estimate sales. *Journal of Economics & Management Strategy*, 1–22. <https://doi.org/https://doi.org/10.1111/jems.12552>
- Navarrete, T., & Borowiecki, K. J. (2016). Changes in cultural consumption: Ethnographic collections in wikipedia. *Cultural Trends*, 25(4), 233–248.
- Neiman, B., & Vavra, J. (2023). The rise of niche consumption. *American Economic Journal: Macroeconomics*, 15(3), 224–264.
- Quan, T. W., & Williams, K. R. (2018). Product variety, across-market demand heterogeneity, and the value of online retail. *The RAND Journal of Economics*, 49(4), 877–913.
- Tan, T. F., Netessine, S., & Hitt, L. (2017). Is tom cruise threatened? an empirical study of the impact of product variety on demand concentration. *Information Systems Research*, 28(3), 643–660.
- Wang, K., & Goldfarb, A. (2017). Can offline stores drive online sales? *Journal of Marketing Research*, 54(5), 706–719.
- Yeo, J. (2022). The rise of niche consumption: A shopping basket similarity approach. *International Economic Journal*, 36(4), 491–509.

A Appendix - Supplementary Figures

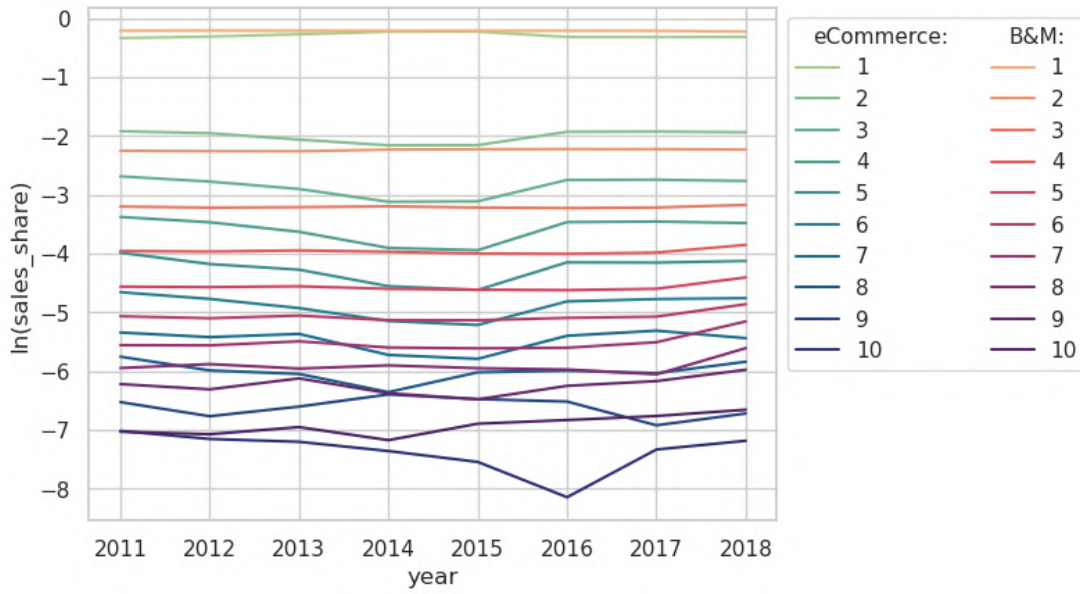


Figure A.1: Log-transformed sum of sales shares of book titles within the respective decile of sales ranks across time. Separated by sales channel. An increase in size of the last deciles (i.e., the long tail), is not apparent.

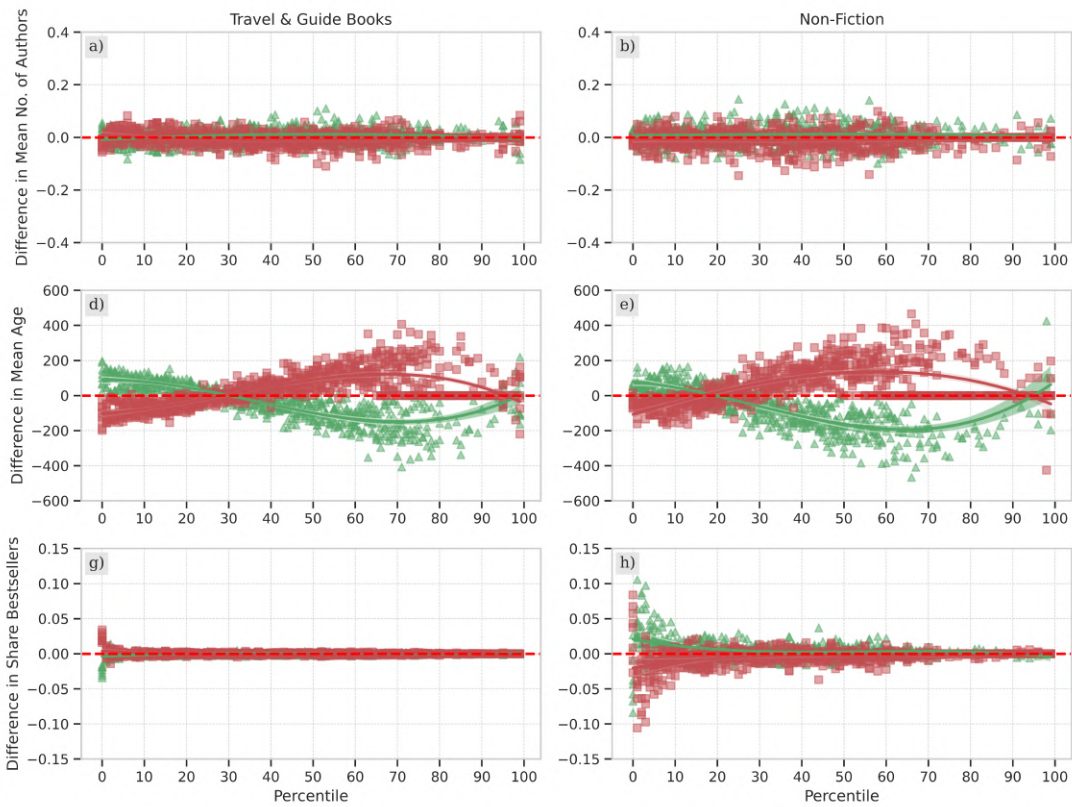


Figure A.2: Difference in books' attributes for different genre-subsets for e-commerce and brick-and-mortar retail overall. The solid lines show the fit of a 3rd degree polynomial linear regression model with 99 % confidence intervals represented by the shaded areas.

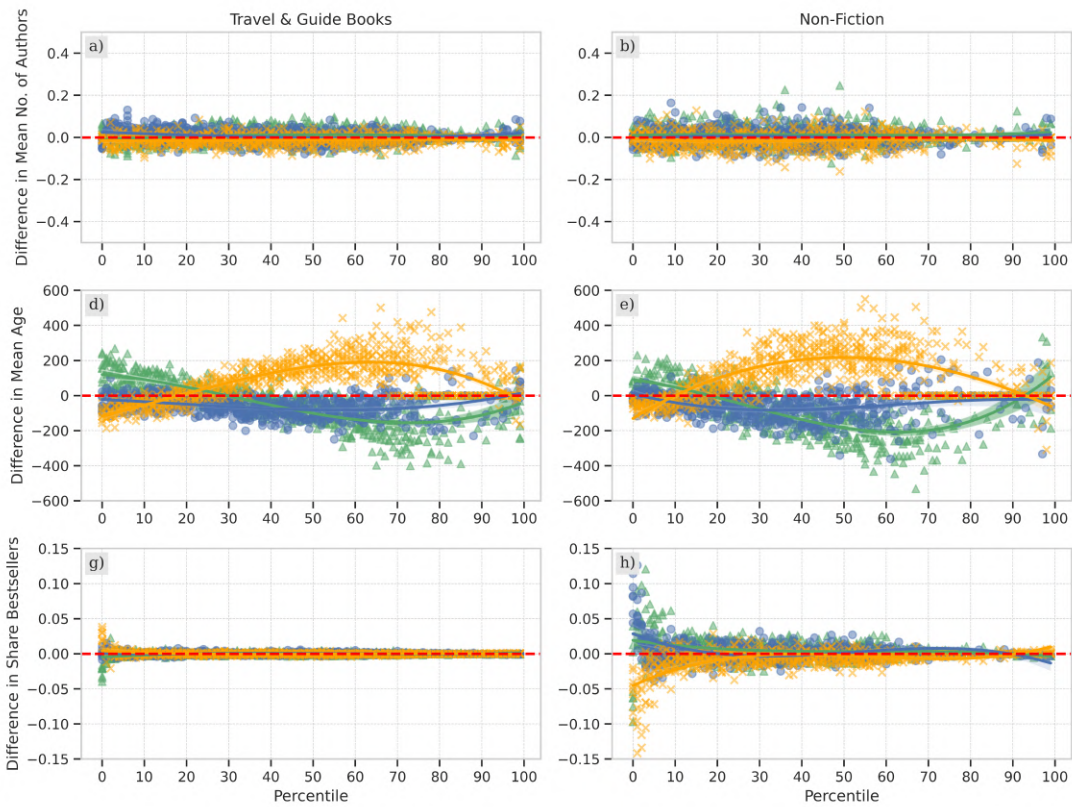


Figure A.3: Difference in books' attributes for different genre-subsets for e-commerce, independent stores, and chain stores. The solid lines show the fit of a 3rd degree polynomial linear regression model with 99 % confidence intervals represented by the shaded areas.