E-book Pricing Under the Agency Model: Lessons from the UK

Maximilian Maurice Gail & Phil-Adrian Klotz

Suggested Citation:
Gail, M. M., & Klotz, P.-A. (2025). E-book Pricing Under the Agency Model: Lessons from the UK. Journal of Industry, Competition and Trade , 25(1), Article 17. https://doi.org/10.1007/s10842-025-00451-y

Journal of Industry, Competition and Trade (2025) 25:17
https://doi.org/10.1007/s10842-025-00451-y

E-book Pricing Under the Agency Model: Lessons from the UK

Maximilian Maurice Gail[1] · **Phil-Adrian Klotz**[2]

Received: 17 May 2024 / Revised: 16 July 2025 / Accepted: 21 July 2025

Abstract
This paper empirically analyzes the relation between the widely used agency model and retail prices of e-books sold in the UK. Using a unique cross-sectional data set of e-book prices for a large number of book titles across all major publishing houses, we exploit cross-genre and cross-publisher variation to examine the interplay between the agency model and e-book prices. Since the genre information is ambiguous and even missing for some titles in our original data set, we also apply a latent Dirichlet allocation (LDA) approach to determine detailed book genres based on the book's descriptions. Using propensity score matching, we find that retail prices for e-books sold under the agency model tend to be systematically lower than book titles with similar characteristics sold under the wholesale model, approximately 20%. This result varies with the exact sales rank of a book and is driven by the so-called *long tail* books. Our results are robust across various regression specifications and double machine learning techniques.

Keywords E-books · Agency agreements · Vertical restraints · Amazon · Propensity score matching

JEL Classification D12 · D22 · L42 · L82 · Z11

1 Introduction

The rise of the Internet—and platform markets more specifically—has accelerated the adoption of so-called agency arrangements. Under this model, suppliers pay retailers sales royalties, and crucially, suppliers determine the final retail prices. On the contrary, in many traditional retail markets, suppliers charge retailers wholesale prices and retailers set final prices to consumers (*wholesale model*). Even though agency arrangements are also used in some of these conventional markets (e.g., newspapers sold at kiosks), this form of a vertical contract is especially prevalent in online markets (e.g., Amazon Marketplace, Apple App

✉ Phil-Adrian Klotz
klotz@dice.hhu.de

[1] Chair for Industrial Organization, Regulation and Antitrust, Department of Economics, Justus-Liebig-University Giessen, Licher Strasse 62, 35394 Giessen, Germany

[2] Düsseldorf Institute for Competition Economics (DICE), Heinrich Heine University Düsseldorf, 40225 Düsseldorf, Germany

Published online: 01 August 2025

Store, eBay Buy It Now). It is also frequently employed in the e-book market, which is the focus of this paper.

In general, books are considered experience goods, as consumers can ascertain their quality only after reading them (Nelson 1970; Reimers and Waldfogel 2021). While some countries, such as Germany, France, or Japan, implement fixed book price (FBP) systems, others, including the UK and the USA, do not. Fixed book prices represent a form of resale price maintenance (RPM) where publishers set retail prices, thereby restricting or eliminating price competition among retailers.[1] The primary motivation for introducing FBP systems is typically to ensure a broad and diverse supply of books, available through a geographically wide network of bookstores.

With the advent of e-books, countries that maintain FBP systems for print books faced the decision of whether to extend this existing legislation to e-books. The applicability of the same cultural policy arguments and legal considerations is debatable, especially given that a geographically wide network of bookstores is irrelevant for e-books. Nevertheless, eight OECD countries with fixed prices for printed books have also extended these fixed prices to e-books, and no country is known to have RPM regulations solely for e-books without corresponding rules for print books (Poort and van Eijk 2017). However, in many countries without fixed prices for e-books (such as the UK), this digital product is partly sold under the agency model (Gilbert 2015), which has similar effects like RPM between a manufacturer and a retailer.[2]

In 2010, Apple, in cooperation with the six largest publishing houses, was the first to adopt the agency model for e-books. This adoption occurred in response to Amazon's aggressive pricing strategy to gain market share. In April 2012, the Department of Justice (DOJ) sued Apple and five of these six publishing houses for conspiring to raise e-book prices by using the agency model in conjunction with most-favored nation (MFN) clauses.[3] Three of the publishers settled shortly after the antitrust case was filed, with the other two settling later the same year. These settlements stipulated that the five publishers could not restrict retailers' ability to set e-book prices for a period of 2 years. Furthermore, the settlements prohibited the use of MFN clauses for 5 years.

Empirical evidence on the price effects of RPM, fixed book prices, and the agency model is scarce. While systematic empirical evidence on RPM is limited to case studies (MacKay and Smith 2017; Ippolito 1991), the only study empirically investigating the agency model's price effect is De los Santos and Wildenbeest (2017). They utilized data on e-book prices for bestselling titles from 2012 and 2013, exploiting the Apple antitrust case as an exogenous shock. Their findings indicate that the agency model, in combination with MFN clauses, led to an average price increase of 8–18% (depending on the retailer).[4] However, De los Santos

[1] Presently, 15 OECD countries have a regulation for fixing the prices of printed books. The fixed prices for printed books typically last 18–24 months after a book has been published.

[2] In particular in combination with most-favored nation (MFN) clauses, the agency model can have economically similar effects where upstream firms control the retail prices (see Sect. 2).

[3] See USA v. Apple Inc., 12 Civ. 2826 (DLC). MFN clauses determine that the retail price set by a publisher of a given product through one retailer must be no higher than the retail price set by this publisher through a competing retailer. Consequently, a MFN clause ensures that if a retailer increases the commission for one publisher, the retail price remains consistent across other retailers. This can incentivize retailers to demand higher fees, ultimately leading to increased retail prices (Johnson 2017).

[4] In a subsequent paper, De los Santos et al. (2021) analyze the reversion from wholesale to agency pricing (following the two-year ban imposed by the Apple case) and report qualitatively similar results for the Amazon platform.

and Wildenbeest (2017) could not disentangle the price effect of the agency model from that of the MFN clauses, as the Apple settlement simultaneously banned both. This limitation represents a key difference from our analysis, as we are able to estimate the price effect of the agency model in the absence of MFN clauses.

The goal of this study is to analyze the price effect of the agency model using a larger and more detailed data set, particularly by not solely incorporating bestselling book titles. We also examine whether similar effects emerge in the absence of an alleged conspiracy, as observed in the Apple case. The internet provides consumers access to a broader range of book titles beyond just the popular ones (bestselling titles), and several authors have demonstrated the importance of so-called *long tail* titles in markets for creative products (e.g., Aguiar and Waldfogel 2018; Brynjolfsson et al. 2003).[5] While a unanimous general definition for these goods is lacking, Brynjolfsson et al. (2011) define the *long tail* as the 50% least sold products. With respect to books, Brynjolfsson et al. (2003) introduced 100, 000 book titles as an absolute cutoff point for long tail, corresponding to the typical inventory and store size of large brick-and-mortar (B&M) bookstores at that time. We adopt this definition and supplement it by also considering 200, 000 book titles, which represents the current inventory and store size of the largest bookstore in London today.[6]

Our cross-sectional data set comprises prices for 12,001 e-books published on Amazon UK between 2010 and 2020. Utilizing data from Amazon ensures an extensive market coverage, given that Amazon accounted for approximately 50% of overall UK book sales (and 87% of UK e-book sales) in 2018.[7] We exploit publisher and book genre variation to estimate the effect of the agency model, beginning with a standard OLS regression approach. We further estimate this effect for different subsets of e-book sales ranks to differentiate between bestselling, middle tail, and long tail books. Subsequently, we apply various robustness checks, notably employing a propensity score matching (PSM) design. Given that the assignment of vertical contract types for e-books was very likely not a random process, PSM helps mitigate selection bias arising from observed book characteristics. This method enables us to match e-books sold under the agency model with comparable e-books sold under the wholesale model based on a range of book characteristics.

The main results from our standard OLS regression approach and the PSM design indicate that e-books sold under the agency model on *Amazon.co.uk* are, on average, approximately 20% cheaper than e-books with very similar characteristics sold under the wholesale model. This effect is heterogeneous, varying with a book's sales rank, i.e., whether a book title is a bestselling or a *long tail* book, and is, according to our findings, almost exclusively driven by *long tail* books. Various robustness checks qualitatively confirm this main finding. This result contradicts the empirical outcome reported by De los Santos and Wildenbeest (2017), but aligns with explanations proposed by the theoretical literature on agency versus wholesale models (Johnson 2020; Foros et al. 2017; Gaudin and White 2014). We discuss the theoretical mechanisms of the agency model and the key differences between our analysis and that of De los Santos and Wildenbeest (2017) in the next section.

The remainder of the paper is structured as follows. Section 2 discusses the related literature. Section 3.1 describes the data, followed by Sect. 3.2 presenting descriptive statistics. Section 3.3 outlines our main estimation strategy. Our primary results are presented in Sect. 4,

[5] For instance, Brynjolfsson et al. (2003) estimated that the benefit consumers obtain from access to *long tail* book titles may be as high as $1.03 billion alone in the year 2000.

[6] The Waterstones Piccadilly Bookshop in London stocks over 200, 000 books at any time (see https://www.timeout.com/london/shopping/waterstones-piccadilly-1, last accessed July 15, 2025).

[7] See Nielsen (2018), "Books & Consumers - UK Industry Standard Report Q4 2018," p. 13.

with extensions provided in Sect. 5. Section 6 details our robustness checks. Finally, 7 concludes and outlines the paper's contribution.

2 Related Literature

Our article contributes to several strands of literature. First and foremost, it relates to studies investigating the competitive effects of the agency model. While the empirical literature on the agency model's economic effects remains rather scarce (apart from the two studies by De los Santos and Wildenbeest 2017 and De los Santos et al. 2021), several recent theoretical papers have analyzed differences in retail prices between the agency and the wholesale model.

Our paper differs from the study by De los Santos and Wildenbeest (2017) in three important ways. First, while De los Santos and Wildenbeest (2017) utilize the court decision in the Apple antitrust case (see Footnote 3) as an exogenous shock in their approach, our findings do not rely on an alleged conspiracy. Second, our empirical analysis incorporates not only bestselling but also *long tail* book titles. We show that the agency effect is heterogeneous with respect to the book sales rank. Third, De los Santos and Wildenbeest (2017) could not isolate the *pure* price effect stemming from agency agreements, as these were implemented in conjunction with MFN clauses during their sample period. In contrast, our analysis can disentangle the effect of the agency model from that of MFN clauses. This is possible because Amazon settled with the EU Commission in 2017, agreeing not to include MFN clauses for any e-book distributed in the European Economic Area (EEA) for the subsequent five years (see AT.40153 E-book MFNs and related matters (Amazon), Decision dated May 4, 2017. We have scraped the Amazon data in 2020).

Our empirical analysis is supported by several theoretical papers analyzing the price effect of the agency model across different market designs. Foros et al. (2017), for instance, model imperfect competition at both the upstream and downstream levels, finding that the agency model's price effect hinges on the relative intensity of competition between these two market levels. Specifically, retail prices will be lower under agency agreements if consumers' price sensitivity between goods (e-books) exceeds their price sensitivity between platforms. This theoretical insight provides an important foundation for our empirical results, particularly given the assumption that competition among publishers should be greater than among retailers—an assumption supported by Amazon's quasi-monopolistic power in the UK e-book market (see Footnote 7). Furthermore, our findings suggest that the negative price effect of the agency model is driven by *long tail* e-book titles, which typically exhibit higher consumer price sensitivity compared to bestselling titles. Nevertheless, Foros et al. (2017) also demonstrate that the agency model can be anti-competitive (leading to higher retail prices) when adopted by platforms on a market-by-market basis. They note, however, that widespread adoption may not always occur due to a prisoner's dilemma among platforms. In this context, MFN clauses can circumvent this prisoner's dilemma by inducing industry-wide adoption of agency pricing, which, in their model, would not otherwise occur, thus rendering them anti-competitive.

Johnson (2020) employs a similar model featuring imperfect competition at upstream and downstream levels, but further assumes that consumers are locked into a single platform. The author finds that when publishers set consumer prices instead of retailers (agency model), prices might be higher initially but lower in later periods. This outcome contrasts with the wholesale model, where retailers strategically set low initial prices to attract and lock in consumers, only to raise them once a sufficient consumers base is established. Lu (2017) utilizes

a bilateral duopoly model with product differentiation in both the upstream and downstream markets to demonstrate that the agency model can benefit consumers—relative to the wholesale model—through lower retail prices due to the elimination of double marginalization. Condorelli et al. (2018) present a theory that endogenizes the choice between agency and wholesale models in an environment where the retailer possesses privileged information about consumer valuations. Their theoretical welfare analysis suggests that the agency model can be welfare enhancing, implying that a ban on agency agreements might decrease welfare. Collectively, these three papers provide robust theoretical foundations that align with our empirical findings.

Another strand of the theoretical literature on agency models focuses on situations where complementary devices are necessary for consuming the main products (e.g., an e-book reader for e-books). Abhishek et al. (2016) find that agency selling is more efficient than the wholesale model, leading to lower retail prices when complementary devices are absent. However, prices under the agency model can sometimes be higher if positive externalities arise from sales of associated products (such as e-readers). Gaudin and White (2014) highlight that a retailer's incentive to set high prices increases when they possess monopolistic control over a complementary device. This was initially relevant in the e-book market when Amazon e-books were exclusively readable on Kindle devices. However, they suggest these effects would have largely dissipated once Amazon released Kindle apps for other platforms in 2009, which is within the relevant period for our empirical analysis.

Our paper also relates to the literature on best price clauses (BPCs) at online platforms. For instance, such BPCs are frequently used by online travel agencies (OTAs), obliging hotels not to offer lower prices on distribution channels other than the respective OTA (narrow BPCs typically cover only the direct channel, wide BPCs extend to other OTAs) (Hunold et al. 2018). Several national competition authorities in Europe have intervened against these clauses, ultimately reaching diverse decisions (Hunold 2017). Most recent empirical literature on this topic indicates that the elimination of (narrow) BPCs primarily incentivized chain hotels to charge the lowest price on the direct channels and hotels also began expanding their room availability on OTAs such as *Booking.com* (Hunold et al. 2018; Mantovani et al. 2021; Ennis et al. 2023).

More generally, our article also relates to the broader literature on RPM. RPM can lead to lower retail prices by internalizing vertical externalities, such as double marginalization (Spengler 1950; Tirole 1988), and RPM regimes can also lead to a larger number of B&M stores compared to regimes with free prices (Dearnley and Feather 2002; Davies et al. 2004). In contrast, Rey and Stiglitz (1988, 1995) show that vertical restraints eliminating intra-brand competition can also serve to mitigate inter-brand competition, thereby becoming anti-competitive. This perspective aligns with Rey and Vergé (2010), who argue that RPM can facilitate industry-wide monopoly pricing, contingent on the extent of potential competition and firms' influence in contractual negotiations. Moreover, several studies find that RPM can be utilized to correct for service externalities (Mathewson and Winter 1984; Perry and Porter 1986; Winter 1993). However, Hunold and Muthers (2017) point out that minimum RPM may distort the allocation of services toward the high-price products of the manufacturer with greater market power.

Finally, our article also contributes to the nascent literature on machine learning (ML) and text mining approaches in economics. Varian (2014) and Athey and Imbens (2019) provide overviews of important ML methods. Wang et al. (2019) employ the *Learning to Place* ML approach to predict book sales, identifying the publishing house as a strong driving factor across all genres. We utilize double machine learning (DML) techniques, similar to the approach of Knaus (2021), who applied DML to examine the impact of children's musical

practice on cognitive skills and school performance. For a broad overview on text mining approaches, see Gentzkow et al. (2019). In this article, we apply a latent Dirichlet allocation (LDA) model to determine book genres by analyzing book descriptions and expert reviews (e.g., Larsen and Thorsrud 2019).

3 Data and Methodology

3.1 Data Set Construction

The data generating process is structured as follows. We scraped the *Amazon.co.uk* webpage for book prices and other book characteristics, with data collection starting in mid-February 2020 taking 2 weeks. To do so, we utilized an a priori list of publishing houses, publishers, and imprints derived from a historical *Sunday Times* bestseller list. This procedure ensures that our sample predominantly comprises books from publishers with a relatively high market share and allows us to cover a broad range of titles.[8]

Our raw data set comprises approximately one million observations, with each containing nearly all available details on the Amazon website, such as prices, formats, descriptions, ratings, and reviews. For each unique book title, there are up to three entries if all formats (hardcover, paperback, and e-book) are available. However, this raw data set includes numerous entries that are either duplicates or irrelevant to our analysis, such as outdated editions or those with incomplete information on specific book characteristics essential for our empirical analysis. Thus, we cleanse the data by retaining only the latest revised version of each title and including only those with available retail prices from our scraped data. Following this cleansing process, our working data set consists of 77, 629 observations (encompassing all three book formats), respectively 47, 161 unique book titles. For our empirical analysis (see Sect. 4), we include only e-book titles for which all explanatory variables (book characteristics) are available (see Table 8 in Appendix 1), resulting in a final working dataset of 12,001 e-book titles.

Our variables of interest are the retail price, which is the price a consumer must pay for a given book, and the treatment variable *Agency*. The latter is coded as one if the Amazon webpage for a book title displays a text field containing "This price was set by the publisher" and zero otherwise.[9] Additionally, our data set contains several control variables for the empirical analysis. These include various book characteristics such as book format, genre, and an e-book size (in KB); book review metrics, including star rating and the number of consumer and expert reviews; information on the author and publisher; and other attributes like the publication date and the recommended retail price (RRP). Table 8 in Appendix 1 provides descriptions for all variables included in our data set.

We also matched the data set obtained from Amazon with a historical *Sunday Times* bestseller list to identify authors who have previously written a bestselling book. This variable is important for our empirical analysis, as a bestselling author's name serves as an important quality signal for the readers.

Each book is a unique product, typically written by a single author and published by a single publisher. Consequently, books are heterogeneous goods, making direct value comparisons between individual titles unfeasible. To appropriately account for this heterogeneity, it is therefore essential to control for book genres in our empirical analysis. However, genre

[8] This bestseller list contains entries from January 2006 until the end of March 2019.

[9] See Appendix Fig. 6 for an example from the Amazon webpage.

information on the Amazon webpage is often ambiguous or entirely unavailable for some titles.[10] Thus, we employ a Latent Dirichlet Allocation (LDA) approach to derive book genres from the descriptions and reviews of the individual books available on the Amazon webpage. This control variable should be appropriate to capture specific effects across genres. We describe this text mining approach in Appendix 2.

3.2 Descriptive Statistics

Our sample comprises 47,161 unique book titles published on *Amazon.co.uk* by Bloomsbury, Faber, Hachette, HarperCollins, Oxford, Pan Macmillan, Penguin Random House, Scholastic, Simon & Schuster, and a group of smaller publishers from 2010 to 2020. While the full data set contains 77,629 observations—reflecting the availability of up to three formats per title—our empirical analysis (detailed in Sect. 4) is based on 12,001 e-book titles. This subset specifically excludes titles with missing book characteristics, as discussed in Sect. 3.1. Although the primary focus of our empirical analysis is on the price of e-books, this section also presents descriptive statistics for hardcover and paperback formats to illustrate the relationships among all three.

Table 9 in Appendix 1 offers descriptive statistics for the variables used in our empirical analysis, summarized by publisher. Beyond e-book retail prices and the RRP, we observe several characteristics for each book title, including its Amazon sales rank, customer ratings, number of customer and expert reviews, and page count. As Table 9 illustrates, e-books from Scholastic exhibit the lowest average retail price, while those e-books from Bloomsbury have the highest mean prices. Additionally, Hachette titles show the lowest average sales rank, and the books published by Simon & Schuster exhibit the highest average number of customer reviews. Most other book characteristics are largely similar across publishers.

Figure 5 in Appendix 1 depicts the frequency distribution of the retail prices for e-books (top), paperbacks (center) and hardcover books (bottom) below £100. Apparently, e-book prices range from £0.25 and £10; paperback prices primarily concentrate in the £10-£20 interval; and hardcover prices are considerably higher. While the distributions of e-books and paperbacks are relatively more compressed, hardcover prices exhibit higher volatility. Finally, all three price distributions exhibit significant mass points at candidate focal points (e.g., £0.49 (e-books), £9.99 (paperback), and £15.99 (hardcover)).

The e-books of individual publishers are sold under different pricing arrangements. Amazon states on its product pages whether the respective publisher has set the price of an e-book. Figures 6 and 7 in Appendix 1 provide examples of this through screenshots of the e-book *Elon Musk: How the Billionaire CEO of SpaceX and Tesla is Shaping our Future* and the e-book *Pulse*. In Fig. 6, the first box on the right-hand side of the Amazon webpage displays the text "price was set by the publisher," indicating an example for the usage of the agency model. In contrast, this information is missing in Fig. 7, meaning Amazon sets the retail price for this e-book, which represents an example of the wholesale model.

Figure 1 visualizes the distribution of e-book retail prices across publishers. Prices are clearly more dispersed for book titles published by Bloomsbury and Oxford, whereas the other major publishers predominantly have books in the range up to £20. The group of smaller publishers has a significantly larger fraction of e-books in the cheapest price range (around £0.49). In Appendix 2, this figure is presented with assigned book genres (see Fig. 13).

Table 1 illustrates the distribution of the pricing arrangements used by individual publishers for the e-books in our sample. In this table, a value of one in the *Agency* column indicates

[10] In Sect. 4, we also present a specification using Amazon's reported genre information as a control variable.

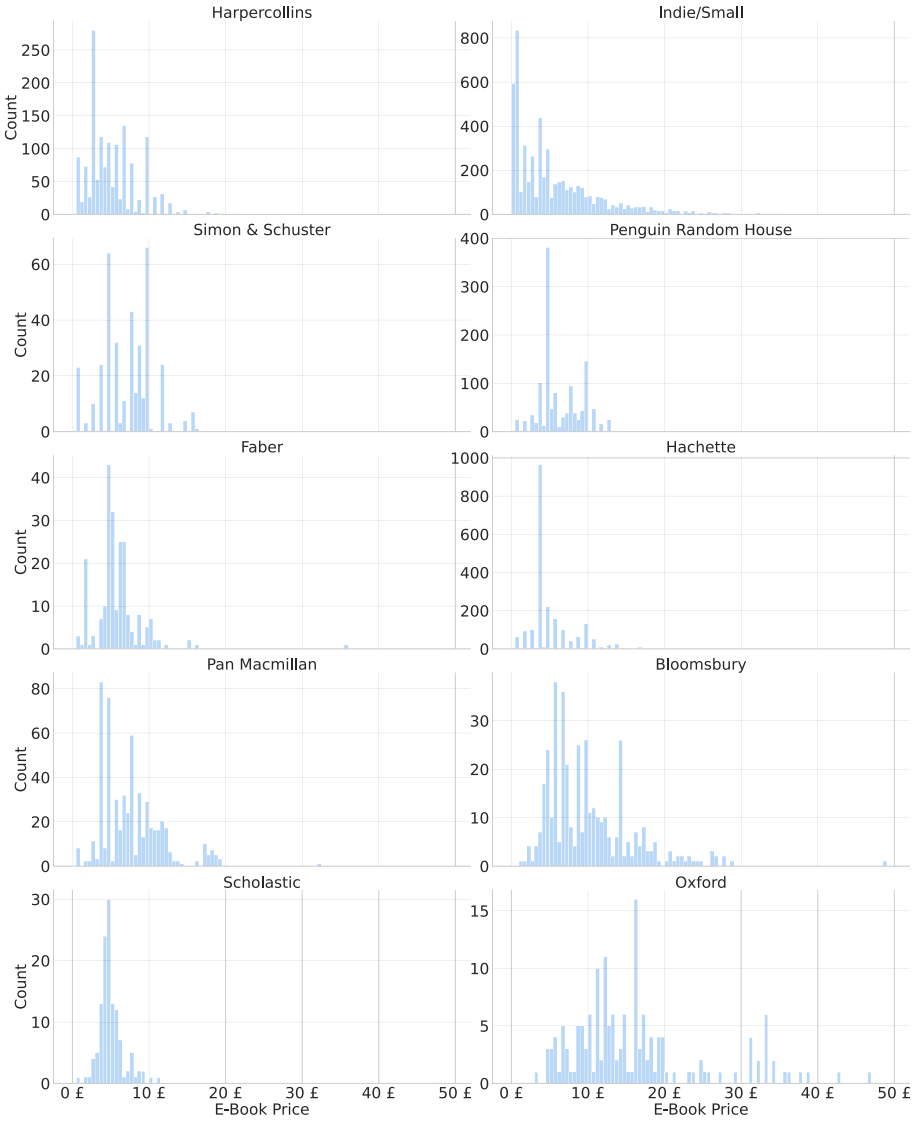

Fig. 1 Prices for e-books grouped by publishers. The interval size for each bar is 1 Pound. For illustration purposes the figures are censored at 50 Pound

that book titles are sold under the agency model, while zero denotes the wholesale model. The table shows that all e-books in our data set published by the large publishing houses Hachette, HarperCollins, Penguin Random House and Simon & Schuster are sold under the agency model on Amazon. Pan Macmillan, however, uses the wholesale model for some of its American imprints' e-books. For e-books in our sample published by Bloomsbury, Faber, Oxford, Scholastic and most smaller publishers, only the wholesale model is used.

Finally, we illustrate the relationship between retail prices for e-books and their book sales rank on Amazon, as depicted in Fig. 2 via a scatter plot with a simple regression line. Notably,

Table 1 Distribution of the agency variable by publishers

Publisher	Agency	Amount	Percentage	Mean price
Bloomsbury	0	397	100%	10.37
Faber	0	223	100%	5.84
Hachette	1	2,073	100%	5.31
Harper Collins	1	1,469	100%	5.42
Small Pub.	0	4,155	77.37%	6.81
	1	1,215	22.63%	2.32
Oxford	0	161	100%	16.71
Pan Macmillan	0	285	50.35%	9.75
	1	281	49.65%	5.81
Penguin Random House	1	1,240	100%	6.56
Scholastic	0	126	100%	5.01
Simon & Schuster	1	376	100%	7.45

a positive relation exists between the retail price and the e-book's sales rank in our sample, as indicated by the positive slope of the regression line. This finding in our sample aligns with the study by Fishwick (2008), who notes that "substantial discounts" (p. 370) have become prominent for bestselling books in the British book market after the abandonment of the Net Book Agreement in 1997.

Regarding book sales rank, it is important to clarify how Amazon determines this metric. According to Amazon, the ranks are internally updated hourly, though they do not appear immediately. Moreover, the rank includes current and *all* prior sales, with higher weight

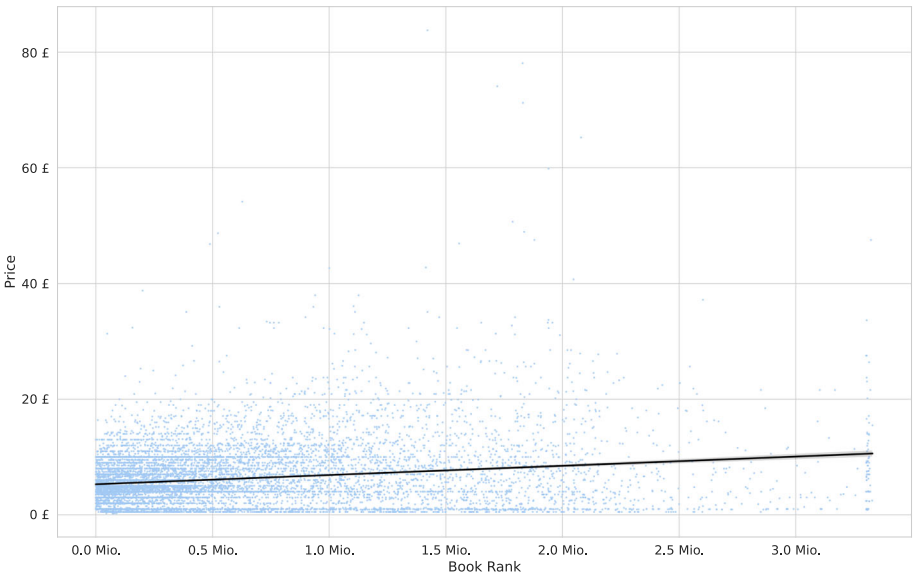

Fig. 2 Relation of Amazon book ranks for e-books and retail prices

given to more recent sales.[11] This information on the definition of book ranks on Amazon will be important for our estimation approach as today's price might be affected by recent or prior sales, but not necessarily the reverse.

3.3 Methodology

The objective of our study is to analyze the relationship between the pricing arrangement and the retail prices of e-books sold on *Amazon.co.uk*. Therefore, we exploit publisher and book genre variation to estimate the price effect of the agency model on e-books. Before turning to the presentation of our estimations, we formalize the hypothesis to be tested. If there were no difference between the two types of vertical contracts (agency and wholesale) with respect to the retail price of an e-book, the publisher's ability to set the final consumer price should (*ceteris paribus*) have no effect on these prices. Hence, the hypothesis to be tested is as follows:

Hypothesis $\boldsymbol{H}_0$. *The retail price of an e-book is independent of the used vertical contract.*

If a positive correlation is found between the agency model and the price of e-books, Hypothesis $\boldsymbol{H}_0$ can be falsified, indicating that e-books sold under the agency model are, on average, more expensive. Conversely, observing a negative correlation would also falsify Hypothesis $\boldsymbol{H}_0$, implying e-books sold under the agency model would be cheaper on average.

In our baseline estimation approach, we employ the standard hedonic modeling approach in the spirit of Rosen (1974), which relies on observing differences in market prices to infer the value or implicit price of underlying characteristics. Thus, we estimate the following *log-log* OLS model with heteroskedasticity-consistent (HC) standard errors[12]:

$$p_i = \alpha_0 + \alpha_1 A_i + \alpha_2 G_i + \alpha_3 R_i + \alpha_4 RRP_i + W\theta + \eta_i. \tag{1}$$

In Eq. (1), the dependent variable p_i is the logarithm of the retail price for e-book i sold on *Amazon.co.uk*. The treatment variable A_i is a dummy variable, which takes the value one if an e-book i is sold under the agency model and zero otherwise. The variable G_i is a categorical variable representing book genres, and R_i is a continuous variable representing the logarithm of e-book sales ranks on Amazon. The variable RRP_i represents the logarithm of the recommended retail price for book title i. All other book-specific covariates are included in the matrix W (see Table 8 in Appendix 1).

However, two potential endogeneity issues may arise in regression Eq. (1), which might distort our estimated coefficients. First and foremost, our treatment variable A_i may be endogenous due to selection bias. Specifically, e-books sold under agency contracts may differ from e-books sold under wholesale contracts in unobserved ways related to their retail prices. Hence, the assignment of an e-book to either the agency or wholesale model was likely not a random process. To mitigate this potential selection bias, we apply a matching procedure in Sect. 5.1.

Second, a bidirectional relationship likely exists between an e-book's rank R_i and its retail price p_i: specifically, beyond rank affecting price, the retail price also influences sales, thereby contributing to rank determination. As already explained, e-book ranks on *Amazon.co.uk* are internally determined by weighted overall sales. Thus, ranks may be driven by the current quantities, but this relationship is ambiguous since the rank also incorporates past sales

[11] See https://kdp.amazon.com/en_US/help/topic/G201648140. (Last accessed: July 15, 2025)

[12] We use the terms heteroskedasticity-consistent standard errors, heteroskedasticity-robust standard errors, and robust standard errors interchangeably; see White (1980) on the concept and Kleiber and Zeileis (2008, Ch. 4.3) for the implementation in R.

quantities. E-book prices can be affected by current and past sales, but the impact of prices on total (current and past) sales is less clear. Nevertheless, we cannot conclusively rule out endogeneity due to a potential reverse causality. To address this potential source of endogeneity between the e-book price and its sales rank, we also present an instrumental variable (IV) approach in Sect. 5.2.

4 OLS Estimation Results

The results of our OLS regression model are outlined in Table 2. We estimate four different specifications in this estimation approach: Column 1 of Table 2 presents a naive OLS regression model without genre effects; column 2 includes LDA genre fixed effects[13]; and column 3 uses the Amazon genre information as a control variable. Column 4 focuses solely on e-books from Pan Macmillan, as this publisher is the only one in our sample (apart from the heterogeneous group of smaller publishers) selling e-books under both vertical contracts types: the agency and wholesale models (see Table 1).

There is a clear and statistically significant negative relationship between the agency pricing arrangement and the retail price of e-books across all four specifications. This provides strong evidence against Hypothesis $\boldsymbol{H}_0$. In terms of magnitude, the effect ranges between 18.29 and 22.04%, depending on the exact specification. For the regression in column 1, an e-book which is sold under the agency model on *Amazon.co.uk* is approximately 18.53% cheaper, on average, than one sold under the wholesale model.[14] Including LDA genre fixed effects (see column 2 in Table 2) increases the agency effect to 18.94% in magnitude, while the effect is slightly lower when incorporating Amazon genre information (see column 3).[15]

The estimated coefficients for the other explanatory variables shown in Table 2 are highly consistent across the first three specifications. The sign of the variable *log sales rank* indicates that e-books with higher sales ranks are sold at higher prices. This observation was previously demonstrated descriptively in Fig. 2 (Sect. 3.2) and confirms the results of Fishwick (2008), who notes that "substantial discounts" became prominent for bestsellers in the UK. However, we discuss potential endogeneity issues concerning this control variable in Sect. 5.2. Moreover, a significant and positive relationship between the RRP of an e-book and its retail price is observable in our results (see variable *log RRP*).[16] We also observe that an e-book's file size (given in KB) has a positive and significant price effect (see variable *log Kindle Size*).[17]

[13] We describe the process to generate book genres using an LDA approach in Appendix 2. Thus, we identify a topic for every e-book title based on the largest probability assigned by the LDA.

[14] To calculate the exact effect of the dummy variable *Agency* on the dependent price variable, the formula $100 \times (e^{\beta} - 1)\%$ must be used (Wooldridge 2013, Section 2.4).

[15] Table 12 in Appendix 2 illustrates that varying the number of topics derived from our LDA text mining approach does not qualitatively change our estimation results.

[16] Theoretically, manufacturers might use the RRP as a focal point to influence retailer pricing decisions (e.g., Motta 2004, Ch. 6.1). An RRP stands out among all possible prices a retailer can choose, potentially coordinating the prices set by individual retailers. Recent empirical studies on this topic have also been conducted (e.g., Faber and Janssen 2019).

[17] Including e-book file size as a control variable allows to account for latent factors that may influence the perceived quality of the e-book (e.g., number of illustrations or image quality), as well as production costs (based on the text length, including typesetting, professional editing, and figure design). Furthermore, file size may indirectly capture the author's opportunity costs, which should be *ceteris paribus* higher for more extensive works.

Table 2 OLS estimation approach. Dependent variable is the logarithm of the retail price for e-books sold on Amazon

	Dependent variable: log Price			
	(1)	(2)	(3)	(4)
Agency	−0.205***	−0.210***	−0.202***	−0.249***
	(0.011)	(0.011)	(0.011)	(0.059)
log sales rank	0.053***	0.059***	0.051***	0.059***
	(0.004)	(0.004)	(0.004)	(0.013)
log Kindle Size	0.045***	0.050***	0.027***	−0.004
	(0.004)	(0.004)	(0.004)	(0.016)
log star rating	0.531***	0.484***	0.436***	0.199
	(0.049)	(0.048)	(0.048)	(0.138)
No. expert reviews	0.015***	0.015***	0.018***	0.028*
	(0.004)	(0.004)	(0.004)	(0.017)
log RRP	1.091***	1.071***	1.080***	0.807***
	(0.011)	(0.012)	(0.012)	(0.045)
Date Retail	0.010***	0.012***	0.008***	0.001
	(0.003)	(0.003)	(0.003)	(0.007)
Bestsellers	0.001***	0.001***	0.001***	0.001
	(0.0004)	(0.0004)	(0.0004)	(0.001)
WeekInChart	0.007***	0.007***	0.007***	0.007
	(0.002)	(0.002)	(0.002)	(0.006)
Constant	−1.856***	−2.204***	−1.745***	−0.876***
	(0.066)	(0.084)	(0.074)	(0.238)
Genre	No	Yes	Amazon Genres	Yes
Inclusion	All	All	All	Pan Macmillan
Robust F Statistic	1285.3022	662.2041	627.4533	67.7247
Observations	12,001	12,001	12,001	566
Adjusted R^2	0.602	0.610	0.615	0.683
Robust standard errors in parentheses	$^{*}p$ <0.1; $^{**}p$ <0.05; $^{***}p$ <0.01			

The results presented in Table 2 also indicate that consumer and expert recommendations seem to drive the price of an e-book, as both the consumer star rating (*log star rating*) and the number of expert reviews (*No. expert reviews*) exhibit a positive and significant effect in our regressions. The time since a book's publication (in years, variable *Date Retail*) also appears to have a positive effect on e-book prices.

Lastly, two covariates related to author quality remain in Table 2. The explanatory variable *WeekInChart* reflects the average number of weeks an author's former bestsellers remained on the *Sunday Times* bestseller charts, and the continuous variable *Bestsellers* represents the historical number of bestselling book titles an author has written. As expected, both variables show a positive and significant effect on e-book retail prices across all three specifications, which can be interpreted as quality signals that increase the price of a book title.[18]

[18] The variables *WeekInChart* and *Bestsellers* are based on a historical *Sunday Times Bestseller* list. The matching process was conducted via Python's Fuzzy Matching.

Column 4 of Table 2, which incorporates only e-books sold by Pan Macmillan, qualitatively confirms our baseline results. Furthermore, focusing solely on e-books from Pan Macmillan in a separate OLS regression can be informative, as it can reduce the selection bias of our treatment variable A_i to some extent (see Sect. 4). In particular, focusing on a single publisher that sells e-books under both vertical contract types allows us to rule out that the agency effect might be confounded with publisher-specific effects (e.g., the relative publisher-platform bargaining power).

The agency effect may vary with the sales rank of a book title, i.e., whether we observe a bestselling title or a *long tail* book (those with higher sales rank). To illustrate this relationship, we re-estimate Eq. (1), starting with a small subset of the data and progressively expanding the sample to include books up to a certain sales rank. For instance, we progressively include up to rank 1000, then up to rank 2000, and so forth, until we reach the largest sales rank. Thus, we obtain a series of estimated agency coefficients, each based on a different e-book subset. Figure 3 illustrates how the coefficients of the agency variable change as higher book sales ranks are gradually included. The black line represents the point estimate, while the shaded area (bounded by red lines) represents the 95% confidence interval. For the lowest range of sales ranks (bestselling books), the coefficient is not statistically different from zero. As book titles from higher sales ranks (*long tail* books) are included, the coefficients become statistically significant and converge to -0.21, consistent with the outcome from column 2 of Table 2. Overall, Fig. 3 implies that the agency effect is strongly dependent on the book sales rank, which can partly explain the divergence between the results of De los Santos and Wildenbeest (2017) (only based on bestselling titles) and our findings.

To emphasize the distinctions among bestsellers (top sales ranks), mid-tail books (medium sales ranks), and *long tail* books (highest sales ranks), we utilize the absolute definition of the 100,000th book rank as the *long tail* threshold (Brynjolfsson et al. 2010). According to Brynjolfsson et al. (2003), a typical large local bookstore stocks between 40,000 to 100,000

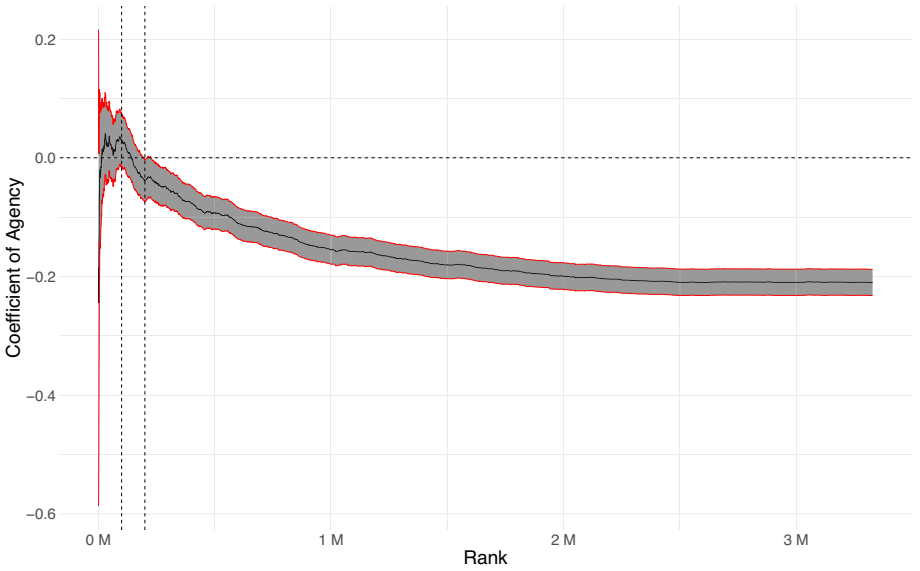

Fig. 3 Results from Eq. (1) if higher sales ranks are gradually included. This starts at sales rank of 1000 comprising approximately 100 observations. The vertical dashed lines denote the book rank of 100,000 and 200,000, representing the potential *long tail* thresholds

unique book titles. Therefore, we classify books with a sales rank exceeding 100,000 as belonging to the *long tail*,[19].

Based on the results presented in Fig. 3, we calculate the mean coefficients by partitioning them at the vertical dashed lines. On average, books in the *long tail* category show an agency coefficient of −15.34 %, while those with ranks below 100,000 exhibit an average effect of approximately 1%.[20]

5 Extensions

As discussed in Sect. 4, the results of our OLS estimation approach may be biased and inconsistent due to endogeneity issues regarding our treatment variable *Agency*. Hence, we apply a matching procedure in Sect. 5.1 to reduce this potential selection bias. Furthermore, we employ an instrumental variable approach in Sect. 5.2 to address the potential endogeneity of our control variable, book sales rank.

5.1 Propensity Score Matching

A simple OLS estimation would allow us to identify the effect of different pricing arrangements if the treatment were random. However, the assignment of e-books to either the agency or wholesale model was likely not a random process, leading to selection bias when using a standard OLS estimation approach. Hence, as a first step, we seek to reduce this selection bias by relying on propensity score matching (Rubin 1977; Rosenbaum and Rubin 1983). Through this, we identify appropriate treated and control e-books through a matching procedure.

In particular, we identify control e-books (not sold under the agency model) that share the same ex-ante probability of being under the agency model as the treated e-books. To implement the PSM, we first estimate a logistic regression to recover the likelihood that an e-book is sold under the agency model based on its observable characteristics and use the predicted values from this estimation to summarize the covariates as a single scalar, the propensity score. Second, we match each e-book sold under the agency model to the most similar e-book sold under the wholesale model, based on this propensity score.

The propensity score e for e-book i is defined as the probability of receiving treatment conditional on the confounding variables X_i: $e(X_i) = Pr(A_i = 1|X_i)$. The identification relies on two key assumptions. The first is the conditional independence assumption, which requires that the outcome variable is independent of the treatment, conditional on the propensity score. This implies that only variables simultaneously influencing both treatment and outcome should be included to ensure unconfoundness. The second assumption, known as the common support (or overlap) assumption, requires that for any given propensity score value, there must be units in both the treatment and the control groups (Rubin 1977). These assumptions are essential for identification, and if both hold, treatment assignment can be considered *strongly ignorable*. Consequently, this allows the analysis to offer unbiased estimates of the treatment effect (Rosenbaum and Rubin 1983; Rubin 1978). The first assumption

[19] In Fig. 3, a vertical dashed line indicates the threshold of 100,000, denoting the point at which the agency coefficient begins to decline and the *long tail* definition applies. Furthermore, for comparative analysis and due to the potential uncertainty regarding the absolute definition, we additionally define the *long tail* using a more recent threshold of 200, 000 (see Footnote 6)

[20] Using a threshold of 200,000 yields comparable findings, with an average coefficient of −15.57% for the *long tail* and approximately zero for ranks below the threshold.

cannot be tested empirically, whereas the second assumption can be evaluated by assessing pre-treatment characteristics and scrutinizing propensity scores once they are estimated. Regarding the latter, we demonstrate below that the common support assumption presents no issues. Concerning the first assumption, we present regression analyses incorporating all covariates previously used from Sect. 4, along with a more limited subset of covariates likely determined prior to the book's sale or the selection of the sales model. This approach is crucial for comparing books under the agency model only with those under the wholesale model that are a highly comparable in their features. We show that estimating the model using PSM can at least substantially decrease bias.

Table 3 Logistic regression with *Agency* as the dependent variable

	Dependent variable: Agency	
	(1)	(2)
log RRP		−0.143***
		(0.041)
log sales rank		−0.175***
		(0.015)
log star rating		−0.184
		(0.134)
No. expert reviews		0.370***
		(0.021)
log kindle size	0.053***	0.119***
	(0.016)	(0.017)
WeekInChart		0.011
		(0.009)
DSeries	−0.422***	−0.363***
	(0.068)	(0.069)
TtS	0.082	0.109
	(0.117)	(0.124)
SReader	−0.099*	−0.139**
	(0.054)	(0.056)
XR	0.535***	0.286***
	(0.047)	(0.050)
Date Retail		−0.005
		(0.010)
Bestsellers		0.019**
		(0.007)
Constant	−0.088	1.471***
	(0.193)	(0.295)
Genre topics	LDA	LDA
Pseudo R2 (McFadden)	0.043	0.103
Adj.Pseudo R2 (McFadden)	0.041	0.099
Observations	12,001	12,001
Log likelihood	−7,891.200	−7,401.500
Robust standard errors in parentheses	$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$	

As a first step of the matching procedure, we estimate the following logit regression:

$$A_i = \alpha + X_i\beta + \eta_i, \tag{2}$$

where A_i is a binary indicator, with $A_i = 1$ for the agency model and $A_i = 0$ for the wholesale model. Hence, $e(X_i) = Pr(A_i = 1|X_i)$ represents the probability that e-book i is sold under the agency model, and X_i is a vector of e-book characteristics (consistent with the explanatory variables in Eq. (1)).

In Table 3, we report the estimates for Eq. (2) from which the propensity scores are estimated.[21] We present results from two PSM specifications: Column 1 includes only a subset of book characteristics as covariates, while column 2 incorporates all explanatory variables introduced in Sect. 4. Specifically, column 1 excludes variables presumed to be realized after the model decision (agency or wholesale) has been made, such as book sales rank, star rating, and expert reviews.

The frequency distributions of the propensity scores based on a 10-nearest-neighbor matching for the treated and control e-books are presented in Fig. 4a and bdepicts the restricted model, while Fig. 4b illustrates the full model. The propensity score distributions of the agency (blue) and wholesale e-books (red) are notably different before the matching procedure (*Unadjusted Sample* in Fig. 4). Following the matching procedure, the propensity score distributions for the two e-book groups are highly similar (see *Adjusted Sample* in Fig. 4), indicating that most wholesale e-books can be matched with agency e-books based on their characteristics. Only a few e-book titles are excluded from the sample after the matching process. Although excluding some of the covariates in this approach clearly alters the distributions between Fig. 4a and b for the treated and untreated e-books, we are able to find a comparable wholesale e-book for nearly every agency e-book in our sample.

In a second step, we use the estimated propensity scores $e(X_i)$ to construct a matched sample comprising only paired e-books to estimate the average treatment effect (ATE) of the agency model on the retail price of e-books. To estimate the ATE, it is essential to impute the counterfactual outcome for each observation i, which is conditional on the applied matching methodology. It is also recommended to incorporate a bias adjustment using the covariates (Abadie and Imbens 2011). Consequently, a regression model is fitted to the outcome variable, specifically the logarithm of prices p, utilizing all previously delineated covariates. This estimation is conducted individually for both the agency and wholesale samples:

$$\begin{aligned} p_i &= W_0\gamma + \varepsilon_i \text{ if } A_i = 0 \\ p_i &= W_1\delta + \mu_i \text{ if } A_i = 1. \end{aligned} \tag{3}$$

From the estimated coefficients, γ and δ, and the values of covariates, W_0 and W_1, we construct bias-adjusted counterfactual values for each missing values $i = 1, \ldots, N$. Specifically, for each observation i with $A_i = 0$, the imputation algorithm follows:

$$\hat{p}_i(A_i = 0) = \begin{cases} p_i & \text{if } A_i = 0 \\ \frac{1}{k}\sum_{j\in N_k(j)}\left[p_j + (W_{0,i} - W_{0,j})\gamma\right] & \text{if } A_i = 1. \end{cases} \tag{4}$$

Similarly, for $A_i = 1$, we obtain the following:

$$\hat{p}_i(A_i = 1) = \begin{cases} \frac{1}{k}\sum_{j\in N_k(j)}\left[p_j + (W_{1,i} - W_{1,j})\delta\right] & \text{if } A_i = 0 \\ p_i & \text{if } A_i = 1. \end{cases} \tag{5}$$

[21] Matching algorithms rely on the R Matching package (Sekhon 2011). Covariate balancing (see Fig. 9) and the propensity score distribution are assessed by using the R package cobalt (Greifer 2024).

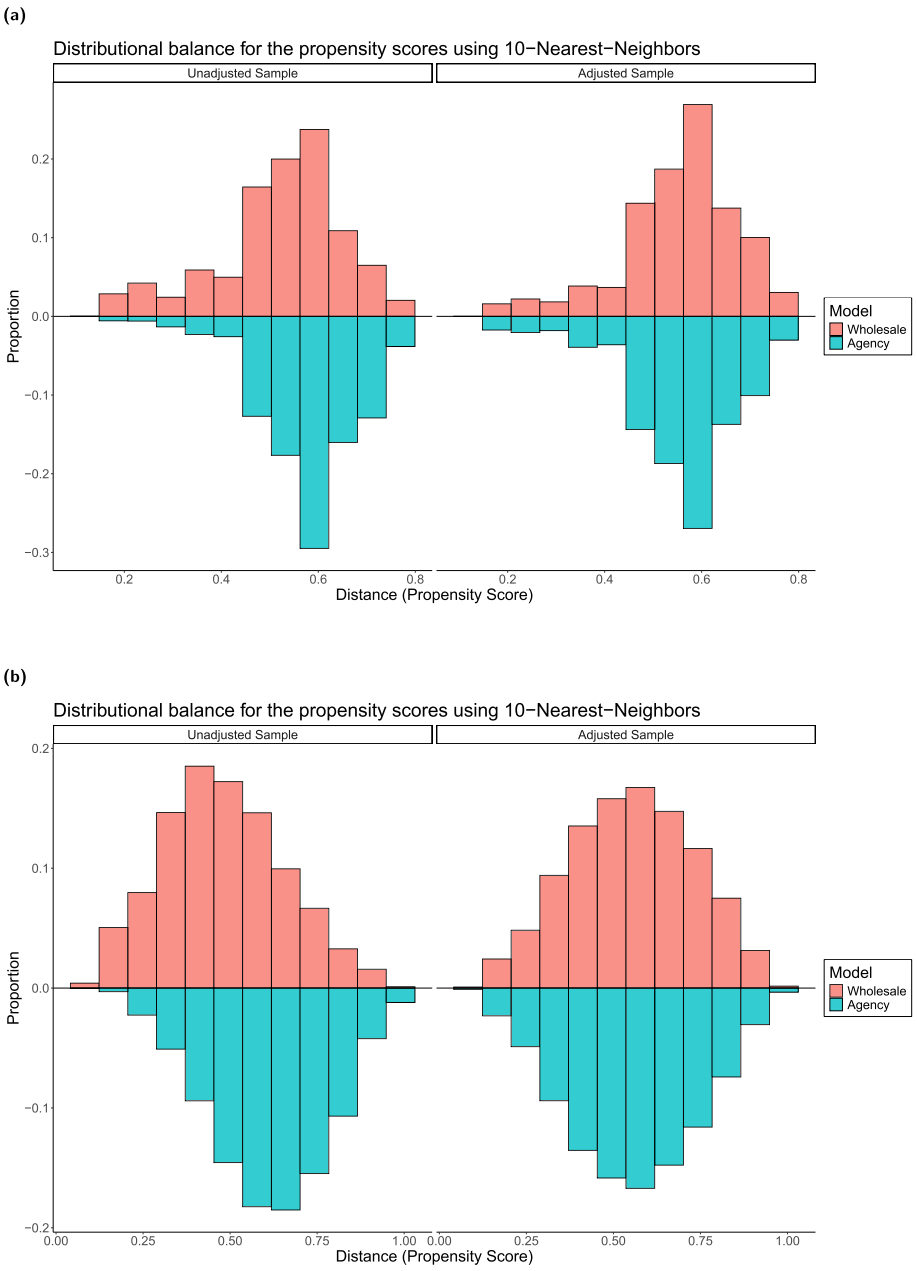

Fig. 4 Frequency distributions of propensity scores (from a logistic model with 10-nearest-neighbors matching): unadjusted (left) and adjusted (right) samples

Table 4 Propensity score matching results for the adjusted subset of covariates, estimated following Eq. (6). Standard errors are Abadie-Imbens heteroskedastic standard errors (except from column 1 of each sub-table, see Abadie and Imbens (2006)). The models in columns 1 to 5 each correspond to a different number of kNN. The caliper radius is set by 0.2 standard deviation of the propensity scores derived from the logistic regression (see Austin 2011, for optimal caliper length)

(a) Restricted model: Based on the propensity scores estimated from the logistic regression in column (1) of Table 3.

	(1)	(2)	(3)	(4)	(5)
$\hat{\tau_{ATE}}$	−0.231	−0.286	−0.304	−0.300	−0.299
Standard error	0.012	0.017	0.016	0.016	0.016
t-statistic	−18.800	−17.212	−18.576	−19.025	−19.071
Number of nearest neighbor (k)	1	1	1	5	10
Replace	No	Yes	Yes	Yes	Yes
Debiased	No	No	Yes	Yes	Yes
Observations	9438	12,000	12,000	12,000	11,999

(b) Full model: Based on the propensity scores estimated from the logistic regression in column 2 of Table 3

	(1)	(2)	(3)	(4)	(5)
$\hat{\tau_{ATE}}$	−0.195	−0.226	−0.207	−0.207	−0.207
Standard error	0.014	0.020	0.019	0.018	0.017
t-statistic	−13.822	−11.104	−10.853	−11.663	−11.854
Number of nearest neighbor (k)	1	1	1	5	10
Replace	No	Yes	Yes	Yes	Yes
Debiased	No	No	Yes	Yes	Yes
Observations	7946	12,001	12,001	11,961	11,941

Here, k denotes the number of matches in our k-nearest-neighbor (kNN) matching, and $N_{k(j)}$ refers to the set of k matches relevant for each observation i. From the imputation algorithms in Eqs. (4) and (5), the final estimator for the ATE is given by (Abadie and Imbens 2011, Section 2.2):

$$\hat{\tau}_{ATE} = \frac{1}{N} \sum_{i}^{N} \left[\hat{p}_i(A_i = 1) - \hat{p}_i(A_i = 0) \right]. \tag{6}$$

Table 4 reports the ATE as estimated from Eq. (6), for the restricted and full model, along with the standard errors, t-statistics, and the count of nearest neighbors employed for each model. In general, we apply nearest neighbor matching methods with a caliper of 0.2 standard deviations, adhering to the common support requirement and addressing the potential non-randomness in the treatment assignment. The agency model continues to show a significant and negative effect on the retail price of e-books. The exact effect ranges from 20.6 to 26.21% (Table 4a) and from 17.7 to 20.2% (Table 4b) depending on the matching procedure used.[22]

[22] Although this may seem like a relatively large effect at first glance, it can be put into perspective by considering an example in absolute terms. The median retail price of an e-book in our data set is £4.99. Assuming that this median e-book is sold under the wholesale model (at £4.99), a 20% negative agency effect in our estimation implies that the same e-books would be sold for £3.99 under the agency pricing model in a counterfactual scenario.

This continues to provide strong evidence against Hypothesis $\boldsymbol{H}_0$. Moreover, the results from the restricted subset of covariates (Table 4a) indicate that excluding variables presumed to be realized post-selection of the sales model does not qualitatively alter our findings, although the coefficient magnitudes are more pronounced.

5.2 IV Estimation

The results of our OLS estimation in Sect. 4 might be biased and inconsistent because our important control variable, book sales rank, may be endogenous. Hence, we employ an instrumental variable approach in this section to address this potential source of endogeneity. We utilize the logarithmized number of customer reviews (*log no. customer reviews*) as an instrument for the book sales rank to avoid inconsistent estimates due to reverse causality.

Our instrumental variable *log no. customer reviews* is highly correlated with our endogenous regressor, book sales rank (relevance condition), but should have no partial effect on the price of an e-book (orthogonality assumption). Customer reviews can enhance the consumer awareness and information quality, thereby influencing the tendency for a consumer to purchase a book. However, the absolute number of customer reviews does not directly affect a consumer's purchasing decision for a book title; only surprisingly positive (negative) reviews can increase (decrease) the consumption of a given good (Reimers and Waldfogel 2021). Hence, the absolute number of customer reviews should also have no partial effect on e-book prices, even though our instrument is highly correlated with the book sales rank (as it is an indicator for past sales).

Following the approach explained above, the linear projection in the first-stage regression of our 2SLS estimation can be formalized as follows:

$$R_i = \beta_0 + \beta_1 A_i + \beta_2 P_i + \beta_3 G_i + \beta_4 P_i \times G_i + \beta_5 RRP_i + \beta_6 CR_i + W\theta + \xi_i. \quad (7)$$

In Eq. (7), the dependent variable R_i refers to the sales rank of book title i on *Amazon.co.uk*. Most covariates used here were introduced in the baseline estimation (Eq. 1). Our instrumental variable *log no. customer reviews* is denoted by CR_i.

The structural equation of our basic model takes the following form:

$$p_i = \gamma_0 + \gamma_1 A_i + \gamma_2 P_i + \gamma_3 G_i + \gamma_4 P_i \times G_i + \gamma_5 RRP_i + \gamma_6 \hat{R}_i + W\theta + \varepsilon_i, \quad (8)$$

where the dependent variable p_i is the logarithm of the retail price of e-book i and the fitted values from the first-stage are captured by $\hat{R}_i$.

The regression results based on Eq. (8) are presented in columns 3 and 4 of Table 5. Column 1 present OLS regression results without including sales rank as a control variable. Column 2 replicates the results of a (naive) OLS estimation from Sect. 4 that includes book sales rank on the right-hand side of the regression. The two IV specifications in Table 5 differ by the type of genre control variable included (LDA-generated or Amazon book genres). The results of our IV approach confirm that e-books sold under the agency model on *Amazon.co.uk* are, on average, significantly cheaper than book titles sold under the wholesale model. Compared to the OLS estimation result in column 2, the estimated coefficients for the variable $Agency$ only differ in magnitude, as we find a negative effect of agency pricing ranging between 17.4 and 18%. Hence, we slightly overestimate the absolute effect of agency pricing when the endogeneity of the variable book sales rank is ignored. The effects of the other explanatory variables also barely differ between the OLS and the IV estimation approaches, although the effect of book sales rank has become larger in the IV regressions.

Table 5 IV estimation results

	Dependent variable: log Price			
	(1)	(2)	(3)	(4)
log sales rank		0.059***	0.090***	0.076***
		(0.004)	(0.008)	(0.007)
Agency	−0.233***	−0.210***	−0.198***	−0.191***
	(0.012)	(0.011)	(0.012)	(0.012)
log Kindle Size	0.053***	0.050***	0.049***	0.026***
	(0.004)	(0.004)	(0.004)	(0.004)
log star rating	0.445***	0.484***	0.505***	0.456***
	(0.048)	(0.048)	(0.048)	(0.048)
No. expert reviews	0.004	0.015***	0.021***	0.023***
	(0.004)	(0.004)	(0.005)	(0.004)
log RRP	1.067***	1.071***	1.073***	1.082***
	(0.012)	(0.012)	(0.012)	(0.012)
Date Retail	0.013***	0.012***	0.011***	0.008***
	(0.003)	(0.003)	(0.003)	(0.003)
Bestsellers	0.001***	0.001***	0.002***	0.002***
	(0.0004)	(0.0004)	(0.0004)	(0.0004)
WeekInChart	0.003	0.007***	0.009***	0.009***
	(0.002)	(0.002)	(0.003)	(0.002)
Constant	−1.421***	−2.204***	−2.612***	−2.076***
	(0.068)	(0.084)	(0.123)	(0.111)
Genre	LDA	LDA	LDA	Amazon Genres
Instrument	No	No	Reviews	Reviews
First stage F-statistic	No	No	3663.8841	3915.4707
Wu-Hausman statistic	No	No	23.179	17.0578
Observations	12,001	12,001	12,001	12,001
Adjusted R^2	0.601	0.610	0.608	0.614
Robust standard errors in parentheses	$^{*}p$ <0.1; $^{**}p$ <0.05; $^{***}p$ <0.01			

6 Robustness Checks

To check the robustness of our results presented in the previous section, we apply two additional estimation approaches. In Sect. 6.1, we utilize data on the paperback format to conduct an alternative OLS estimation procedure. Subsequently, we present a double machine learning (DML) approach in Sect. 6.2.

6.1 Format Estimation

The coefficient for the agency variable in our OLS estimation approach might be biased since we compare different book titles with each other. An ideal analysis would compare the same title offered to similar consumers under two separate vertical contracts. However,

a specific e-book title is sold under either the agency or the wholesale model, rendering the counterfactual scenario unobservable. As an alternative, we conduct a robustness check wherein we compare price differences between digital and paperback formats for the same book title. Since print books are generally sold under wholesale contracts in the UK, this allows us to identify the price effect of the agency model on e-book prices (by also controlling for format differences).

To implement this robustness analysis, we retain only e-books sold under the agency model that also have an equivalent paperback version in our sample, which reduces our data set to 7446 observations. Then, we estimate the following OLS model:

$$p_{i,j} = \alpha_0 + \alpha_1 A_{i,j} + \alpha_2 F_{i,j} + W\theta + \epsilon_{i,j}. \tag{9}$$

The dependent variable $p_{i,j}$ in Eq. (9) is the logarithm of the retail price for a book title i in format j. Again, the treatment variable $A_{i,j}$ is a dummy variable, equal to one if a book is sold under the agency model and zero for the wholesale contract. The format fixed effect $F_{i,j}$ is equal to one for the paperback version and zero for the digital version. All other covariates are captured by the matrix W.

The results of Eq. (9) are reported in Table 6, controlling for genre fixed effects (column 1), publisher fixed effects (column 2), and both types of fixed effects (column 3). Column 4 estimates Eq. (1) by incorporating only the e-book format so that all the observations for the

Table 6 Robustness check with paperback books (log-log OLS). Dependent variable is the logarithm of the retail price for digital and paperback versions of book titles sold on Amazon

	Dependent variable: log Price			
	(1)	(2)	(3)	(4)
Agency	−0.174***	−0.261***	−0.236***	−0.166***
	(0.017)	(0.021)	(0.021)	(0.010)
Paperback	0.485***	0.439***	0.447***	
	(0.014)	(0.015)	(0.015)	
log star rating	0.219***	0.215***	0.182***	0.281***
	(0.045)	(0.046)	(0.045)	(0.043)
No. expert reviews	0.038***	0.032***	0.031***	0.039***
	(0.005)	(0.006)	(0.006)	(0.004)
log RRP	0.888***	0.931***	0.888***	0.839***
	(0.014)	(0.013)	(0.014)	(0.011)
Date Retail	0.004	0.005	0.003	0.014***
	(0.004)	(0.004)	(0.004)	(0.003)
Bestsellers	−0.0002	−0.001***	−0.001**	−0.00002
	(0.0002)	(0.0002)	(0.0002)	(0.0003)
Constant	−0.479***	−0.398***	−0.257***	−0.462***
	(0.058)	(0.037)	(0.059)	(0.064)
Publisher	No	Yes	Yes	No
Genre	Yes	No	Yes	Yes
Robust F Statistic	497.1137	627.141	381.3814	29.6844
Observations	7,446	7,446	7,446	3,723
Adjusted R^2	0.606	0.609	0.622	0.479
Robust standard errors in parentheses	$^{*}p$ <0.1; $^{**}p$ <0.05; $^{***}p$ <0.01			

paperback format are excluded (see the number of observations and the empty cell for the paperback coefficient in column 4). Comparing the estimation results to those reported in Table 4b, we observe that the agency effect is highly similar, although its magnitude is partly smaller here (ranging between 15.3 and 23%). The estimated coefficients for the variable *Format* imply that the paperback versions are significantly more expensive than the digital ones, confirming our descriptive statistics in Sect. 3.2 (see Fig. 5). The coefficients of the remaining covariates in Table 6 are largely consistent with our main estimations in Sect. 4.

The results of a slightly modified regression model based on Eq. (9) are presented in Table 10 of Appendix 1. Here, the dependent variable is the difference in absolute prices between digital and paperback formats for a given book title. This makes the interpretation for the agency effect more intuitive, although the treatment effect must now be understood as an absolute effect (in British pounds) rather than a percentage.[23] In Table 10, the constant of the regression line represents the mean price difference between e-books and paperbacks, indicating that the print format is more expensive than the digital format, on average. The significant negative coefficient for the variable *Agency* implies that this price difference is more pronounced for agency compared to wholesale e-books. This further suggests that e-books sold under the agency model are cheaper than e-books sold under the wholesale model, which confirms our main results in Sect. 4.

6.2 Double Machine Learning (DML)

Beyond established approaches in the standard econometric analysis for causal inference, further methods exist. We have already applied standard econometric methods as the OLS estimation and the propensity score matching to deal with econometric issues. Recent advances in machine learning approaches also provide an expanded toolbox for empirical analyses in economics (e.g., for a broad overview, see Athey and Imbens 2019; Athey 2018).

Recent developments in the machine learning literature have yielded many approaches, such as the DML technique, that enable addressing common econometric issues (including confounding variables or variable selection) through the use of cross-validation and non-parametric models. This technique also allows non-standard modeling of relationships between variables, unlike conventional approaches where independent variables often assume specific, linear or quadratic effects on the dependent variable. It permits the use of any arbitrary machine learning technique, which relies on algorithms to find a suitable model for some chosen score functions, such as the mean squared error.

The reason for relying on further non-parametric/semi-parametric regressions is to circumvent the imposition of a specific model structure. The algorithms of the machine learning models will select the best-fitting model under given restrictions or parametrization. We then employ regularized linear regression techniques, such as the least absolute shrinkage and selection operator (Lasso), or regression trees/forests. These methods facilitate comparing our standard econometric approaches with models that are capable to disregarding irrelevant variables or incorporating non-linear effects (Athey and Imbens 2019). Moreover, this methods can assist in a similar manner to propensity score matching in addressing the underlying selection issue (e.g., Lee et al. 2010; Westreich et al. 2010; Knaus 2021).

Therefore, we apply DML techniques and compare the results to our previous estimations to provide additional robustness checks. From a predictive perspective, the estimations are split into three distinct regression models, as proposed by Chernozhukov et al. (2018). These

[23] Appendix 1 Table 11 also presents the results in percentages, derived using log ratios (differences in logs) to compare digital and paperback formats.

models utilize the DML framework to address challenges such as high-dimensional variables, non-parametric functional forms, robust variable selection, or unobserved confounding variables. While the DML framework can be extended to handle endogeneity arising from unobserved confounding variables (the latter typically requiring instrumental variable methods), our current application focuses on handling observed covariates. The analysis is based on the *DML-Conditional-Average-Treatment-Effect-Estimator*.[24]

The equation system we estimate takes the following general form (dropping the individual index i for each book), which is based on the partial linear regression model of Robinson (1988):

$$\begin{aligned} p &= \theta A + q(W) + \varepsilon \\ A &= f(W) + \eta \\ s.t.\ \mathbf{E}[\varepsilon|W] &= \mathbf{E}[\eta|W] = \mathbf{E}[\varepsilon \cdot \eta|W] = 0, \end{aligned} \tag{10}$$

where p denotes the price of an e-book, which depends on the agency dummy variable A and $q(W)$ is a function of covariates W. The agency dummy variable A is modeled by a function $f(W)$ of these covariates W (similar to the logit specification in Eq. (2)). The variables η and ε represent stochastic error terms. These conditional expectation functions can then be estimated by using non-parametric regressions:

$$\begin{aligned} q(W) &= \mathbf{E}[p|W] \\ f(W) &= \mathbf{E}[A|W]. \end{aligned} \tag{11}$$

Next, the residuals for the price, $\tilde{p}$, and the residuals of the agency dummy variable, $\tilde{A}$, are computed by subtracting the fitted values (obtained from the regression tasks in (11)) from the actual price p and the actual agency dummy variable A:

$$\begin{aligned} \tilde{p} &= p - q(W) \\ \tilde{A} &= A - f(W) \end{aligned} \tag{12}$$

Finally, we use these residuals (for prices $\tilde{p}$ and the agency dummy variable $\tilde{A}$) to estimate a linear treatment effect θ which is unbiased based under the assumptions of Chernozhukov et al. (2018):

$$\tilde{p} = \theta \tilde{A} + \epsilon. \tag{13}$$

In Table 7, the column *Model* represents the applied functional form. The first entry in this column refers to the functional form used for computing and predicting $\tilde{p}$, and the second for classifying $\tilde{A}$. Specifically, *Lin-Logit* relies on OLS and a logistic regression; *Lin-Lasso* relies on an OLS and a logistic regression including L_1 penalty (called Lasso); *Lasso-ElasticNet* employs a Lasso and a logistic regression with a combination of L_1 and L_2 penalties (*Elastic Net)*; *Lasso-RFC* refers to a Lasso regression and a random forest classifier; *RFR-ElasticNet* combines a random forest and an *Elastic Net*; *RFR-RFC* utilizes a random forest for both stages; and *XGBoost* refers to Extreme Gradient Boost for both stages.[25]

The column *Score* in Table 7 displays the mean squared error from the final stage. In the final stage, a simple linear regression is used to estimate the conditional average treatment

[24] For the estimation implementation, we follow the Python module *econml* provided by Battocchi et al. (2019) (see also https://econml.azurewebsites.net/spec/api.html#linear-in-treatment-cate-estimators, last accessed, July 15, 2025).

[25] While many other estimation techniques are possible, these are sufficient to highlight the stability of our results.

Table 7 Double machine learning approach. The dependent variable is the logarithm of the e-book retail price and the treatment variable is *Agency*. The price effect of the agency arrangement for each model is given in the column *Agency* and represents the ATE. Column *Score* refers to the mean squared error. Note: The data has been mean centered with unit variance, as Lasso requires this normalization

	Agency	Std. Error	p-value	Score	Perc. Change
Lin-Logit	−0.2138	0.0115	0.0000	0.3263	−19.2519
Lin-Lasso	−0.2138	0.0115	0.0000	0.3263	−19.2485
Lasso-ElasticNet	−0.2137	0.0115	0.0000	0.3264	−19.2443
Lasso-RFC	−0.2417	0.0132	0.0000	0.3266	−21.4719
RFR-ElasticNet	−0.1714	0.0094	0.0000	0.2160	−15.7488
RFR-RFC	−0.2057	0.0114	0.0000	0.2153	−18.5962
XGBoost	−0.1688	0.0119	0.0000	0.2314	−15.5280

effect. The hyperparameters for each model are selected from a reasonable set and then we use $3-5$ cross-fold validation within Python's Sklearn GridSearch. Additionally, an outer 5-fold splitting is also applied in each estimation. The presented results outline the best estimate (lowest score) for each model class.

The point estimates of the individual regression models are given in the column *Agency*, and the relative percentage changes (in relation to the intercept) are presented in the column *Perc. Change* of Table 7.[26] Overall, the DML techniques confirm the results of our main estimations presented in Sect. 4, reaffirm strong evidence against Hypothesis $\boldsymbol{H}_0$, and demonstrate the robustness of our regressions, even when employing more flexible methods. For instance, e-books sold under the agency model are, on average, 18.6 % cheaper than digital books sold under the wholesale model when using the regression model *RFR-RFC*.

7 Conclusion

In this paper, we provide evidence that e-books sold under the agency model on *Amazon.co.uk* are, on average, significantly cheaper than e-books sold under the wholesale model. Our results are based on a unique data set encompassing numerous characteristics of an e-book. To quantify the relationship between the retail price of an e-book and pricing arrangement used, we rely on classical econometric techniques as well as on newer methods, such as the DML approach. We find a robust and statistically significant effect that e-books sold under the agency model are approximately 20% cheaper than e-books books sold under the wholesale model. The exact effect depends on the sales rank of a book, that is, whether a book title is a bestselling or a *long tail* book. Moreover, our findings indicate that this effect is driven almost exclusively by *long tail* books.

The results of our empirical analysis align with many theoretical papers analyzing the price effect of the agency model (cf. Sect. 2). These theoretical papers find that retail prices for e-books sold under the agency model are lower due to the elimination of double marginalization (Lu 2017; Condorelli et al. 2018), a lock-in effect exploited by retailers (Johnson 2020), or because agency selling is more efficient than the wholesale model and leads to lower retail prices (Abhishek et al. 2016). Additionally, our results align with the theoretical analysis by

[26] The exact percentage change can be calculated using the formula $100 \times (e^{Agency} - 1)\%$ (Wooldridge 2013, Section 2.4).

Foros et al. (2017) if one assumes that competition is greater among publishers than among retailers, which most likely is the case given Amazon's quasi-monopolistic power in the UK e-book market (see Footnote 7). Our finding that the negative price effect of the agency model is driven by *long tail* e-book titles, which exhibit higher consumer price sensitivity compared to bestselling titles, is theoretically supported by Foros et al. (2017).

To the best of our knowledge, this paper is the first empirical analysis to estimate the price effect of the agency model for e-books in the absence of MFN clauses and by incorporating not only bestselling titles but also *long tail* book titles. Besides, we also apply an LDA approach to determine book genres. Nevertheless, a limitation of our approach is its reliance on cross-sectional data instead of panel data, which prevents us from controlling for dynamic effects on e-book retail prices. Moreover, we include only one online platform, namely Amazon, instead of comparing various online retailers. Although Amazon has a relatively high market share for e-books in the UK (see Footnote 7), competition between the individual retailers very likely also affects e-book prices.

The dynamic effect of the agency model on e-book prices, including bestselling as well as *long tail* book titles, remains an open question. Future research should concentrate on panel data to address such dynamic effects of different vertical contracts. Furthermore, other online platforms selling e-books should also be included in such an analysis to identify effects between the online retailers. Finally, the long-run effect of the agency model on consumer welfare presents an interesting research area. Consumer welfare depends not only on e-book prices, but also on other factors such as the number, variety, and quality of book titles written and published.

Appendix 1: Additional figures and tables

Table 8 Relevant variables per book title and the information content they provide

Variables	Information
Price	Retail price from the upper right *Buy-Box*
Format	Hardcover, paperback, Kindle
Star rating	Average rating normalized to be between 0 and 1
No. customer reviews	Number of consumer reviews
No. expert reviews	Number of expert reviews on Amazon
DSeries	Dummy variable whether book is part of a series
Description and reviews	Detailed text-information on the book and by different reviewers
Genre	Constructed by LDA from the descriptions and reviews (see Appendix 2)
RRP	Recommended retail price which is the print RRP. For Kindle it is either related to the hardcover or paperback RRP
Agency	Dummy variable to be one if the price was set by the publisher and zero otherwise. Only possible for e-books
Seller	Sample is restricted to be sold by Amazon
Author	Information on the author of a book
Title	Information on the title of a book
Kindle_Size	Kindle file size (in KB)

Table 8 continued

Variables	Information
Publisher	Name of the publisher. We have different levels of aggregation (Imprint, Publisher, Publishing House)
Amazon rank	Uncategorized Amazon bestseller rank for either print books or e-books
Bestsellers	Number of bestsellers in the Sunday Times Bestseller List conditional on the Author's name
WeekInChart	Average number of weeks in the bestseller charts conditional on the Author's name
Identifier	Aggregation of ASINs to verify the books
Date Retail	Period of time since the publication of a book title (in years)
TtS	Dummy variable whether Text-to-Speech is enabled for an e-book
SReader	Dummy variable whether Screen Reader is supported for an e-book
XR	Dummy variable whether enhanced typesetting is enabled for an e-book

Table 9 Summary statistics

		Total	Bloomsbury	Faber	Hachette	Harpercollins	Indie/Small	Oxford	Pan Macmillan	Penguin Random House	Scholastic	Simon & Schuster
Retail Price	mean	6.18	10.37	5.84	5.31	5.42	5.80	16.71	7.79	6.56	5.01	7.45
	std	5.17	7.02	3.21	2.86	3.20	6.07	10.12	3.86	2.70	1.58	3.27
Sales Rank	mean	567929.29	554409.87	326982.37	262402.04	427905.75	778429.39	937724.43	497575.36	365604.06	597692.75	555114.93
	std	650399.53	594258.21	432500.62	383524.36	514257.98	728316.54	653612.74	551601.82	533573.80	655036.93	649582.93
Star Rating	mean	0.88	0.90	0.85	0.89	0.89	0.86	0.90	0.89	0.88	0.93	0.90
	std	0.10	0.10	0.13	0.08	0.09	0.12	0.10	0.10	0.08	0.07	0.08
No. Customer Reviews	mean	97.92	51.10	60.80	119.85	106.88	80.50	17.19	124.70	134.59	76.33	142.89
	std	171.91	110.04	113.83	178.01	168.94	161.48	42.95	210.71	195.05	138.88	210.61
Pages	mean	320.98	292.57	280.60	388.78	342.59	286.86	412.85	336.34	319.01	231.84	367.91
	std	841.07	124.35	191.93	1949.17	330.58	212.83	207.28	121.63	140.63	113.11	256.70
Kindle Size	mean	11470.46	13078.55	2849.00	14134.92	8288.26	8626.52	9995.83	12118.73	20466.37	35684.97	15118.97
	std	35870.99	27346.77	10529.46	46813.98	28600.24	30096.99	16416.27	37011.08	48498.42	42097.93	27885.95
RRP	mean	12.98	15.50	10.28	13.07	12.04	12.08	35.29	13.61	14.59	8.31	13.85
	std	9.26	9.76	4.88	5.32	5.90	10.22	30.15	5.68	5.91	3.15	5.29
Date Retail	mean	1.96	1.64	2.41	1.81	2.09	1.93	2.97	1.66	2.18	2.11	2.13
	std	2.06	1.76	2.62	2.06	2.36	1.91	1.81	1.78	2.25	2.11	2.29
No. Expert Reviews	mean	1.60	1.74	2.43	2.33	1.25	1.34	1.14	2.24	1.71	0.65	1.35
	std	1.40	0.96	1.24	1.06	1.12	1.51	0.77	1.57	1.40	0.85	1.04

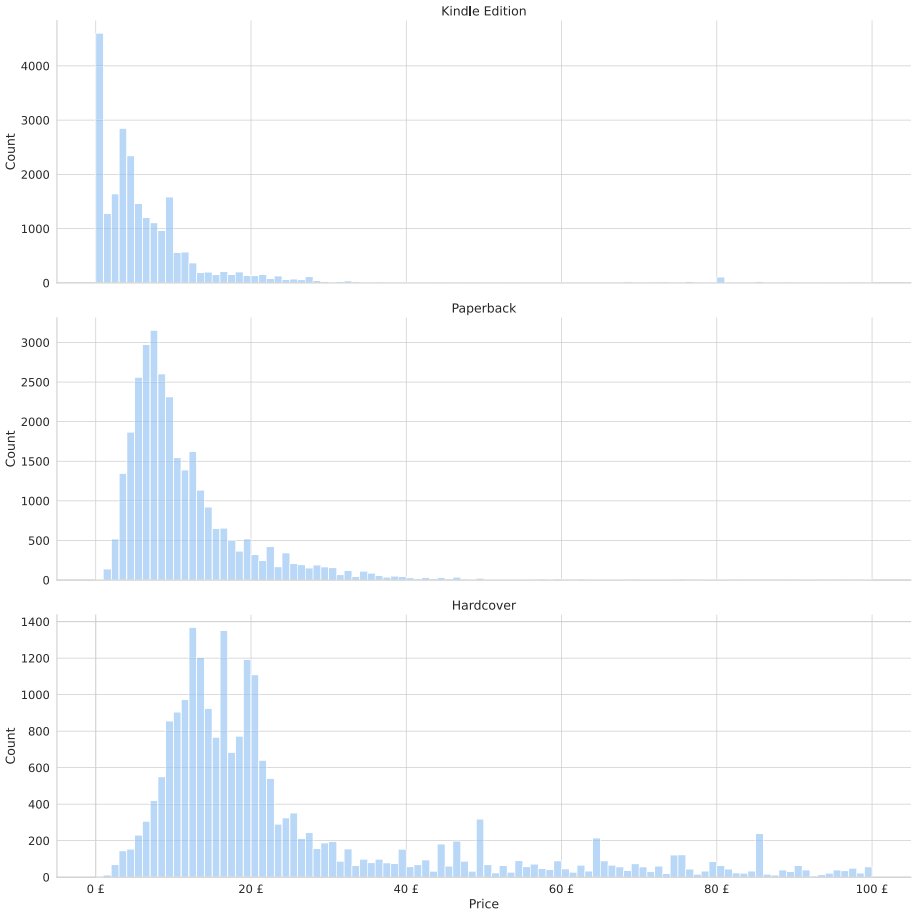

Fig. 5 Distribution of retail prices by book format

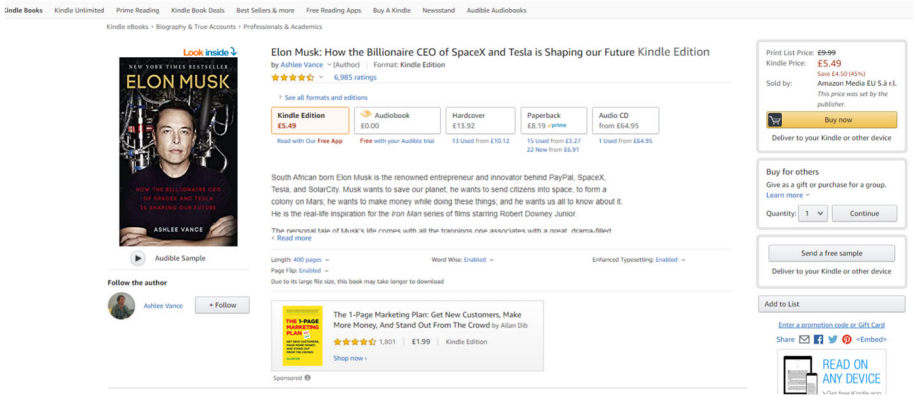

Fig. 6 Screenshot of *Elon Musk: How the Billionaire CEO of SpaceX and Tesla is Shaping our Future* (*Amazon.co.uk*)

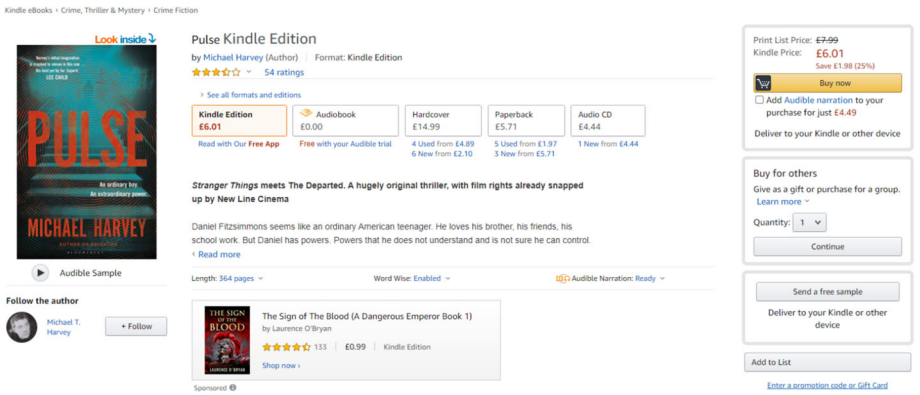

Fig. 7 Screenshot of *Pulse* (*Amazon.co.uk*)

Table 10 Robustness check with paperback books. Dependent variable is the difference in the retail price between digital and paperback versions of book titles sold on Amazon

	Dependent variable: Kindle Price - Paperback Price		
	(1)	(2)	(3)
Agency	−0.385***	−0.265**	−0.436***
	(0.111)	(0.114)	(0.113)
log star rating	1.811***	1.597**	1.039
	(0.671)	(0.667)	(0.676)
No. expert reviews	−0.012	−0.011	0.024
	(0.042)	(0.042)	(0.039)
log RRP	0.378***	0.330**	0.714***
	(0.140)	(0.158)	(0.151)
Date Retail	−0.038	−0.047*	−0.038
	(0.029)	(0.029)	(0.028)
Bestsellers	0.007***	0.007***	0.008***
	(0.002)	(0.002)	(0.002)
Constant	−3.218***	−3.333***	−4.461***
	(0.381)	(0.713)	(0.463)
Genre	No	LDA	Amazon
Robust F Statistic	8.0357	6.7923	19.4451
Observations	3,723	3,723	3,723
Adjusted R^2	0.016	0.032	0.059
Robust standard errors in parentheses	$^{*}p$ <0.1; $^{**}p$ <0.05; $^{***}p$ <0.01		

Table 11 Robustness check with paperback books. Dependent variable is the difference in the log retail price between digital and paperback versions of book titles sold on Amazon

	Dependent variable: log(Kindle Price) - log(Paperback Price)		
	(1)	(2)	(3)
Agency	−0.099***	−0.077***	−0.106***
	(0.020)	(0.020)	(0.020)
log star rating	0.300***	0.243**	0.142
	(0.104)	(0.102)	(0.103)
No. expert reviews	0.011	0.013	0.020***
	(0.008)	(0.008)	(0.007)
log RRP	0.316***	0.309***	0.369***
	(0.021)	(0.023)	(0.023)
Date Retail	0.002	−0.001	0.001
	(0.005)	(0.005)	(0.005)
Bestsellers	0.001**	0.001**	0.002***
	(0.001)	(0.001)	(0.001)
Constant	−1.195***	−1.279***	−1.373***
	(0.061)	(0.126)	(0.072)
Genre	No	LDA	Amazon
Robust F Statistic	38.1151	27.3776	25.7949
Observations	3,723	3,723	3,723
Adjusted R^2	0.092	0.117	0.141
Robust standard errors in parentheses	$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$		

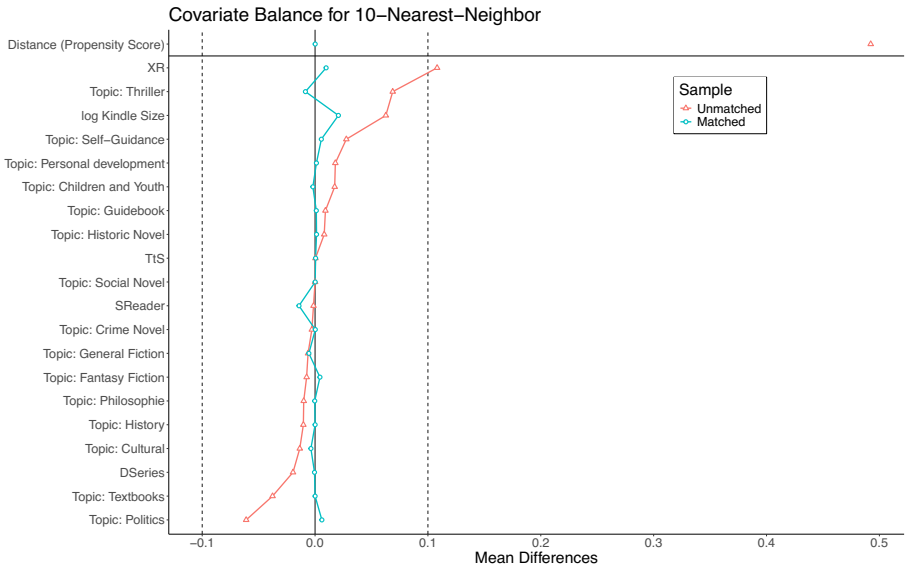

Fig. 8 Covariate balance for 10-nearest-neighbors (see column 5 of Table 4a in Sect. 5.1). The x-axis represents standardized mean differences, while the y-axis lists the covariates introduced in Sect. 3.3. The figure compares the balance between unmatched (red line) and matched (blue line) samples. The goal of matching is to reduce imbalances between the treatment and control groups, a goal reflected by the blue line moving closer to zero relative to the red line. Dashed vertical lines indicate a common balance thresholds (e.g., 0.1), which assist in assessing whether covariates are sufficiently balanced post-matching. The results suggest that the matching procedure significantly reduces imbalances for all covariates

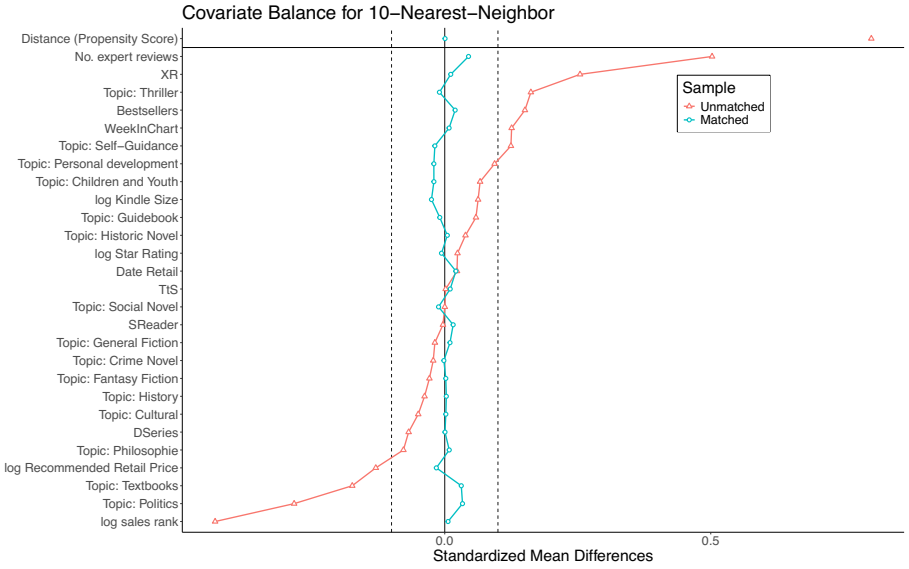

Fig. 9 Covariate balance for 10-nearest-neighbors (see column 5 of Table 4b in Sect. 5.1). Refer to Fig. 8 for interpretation details

Appendix 2: Latent Dirichlet allocation (LDA)

In the recent past, new technologies have enabled the use of text as data, thereby, integrating it as an input to economic research. Text data, which is inherently high-dimensional, can capture relevant economic concepts not covered by *hard* economic data. In the recent years, there has been an explosion of empirical economics research using text as data (e.g., see Larsen and Thorsrud (2019) for a latent Dirichlet allocation (LDA) approach or Lenz and Winker (2020) for paragraph vector topic modeling). We have decided to use an LDA approach to generate book genres and to assign each single book title from our data set into one of these genres. Such a text mining approach is necessary because on the Amazon webpage, genre information is ambiguous or even unavailable for some book titles. For this purpose, we use the descriptions and expert reviews from individual books in our data set as text data input. We further rely on natural language processing (NLP) to extract the relevant information.

We apply several Python modules to clean and prepare the raw data set.[27] This process involves removing common words and surnames, eliminate stop words, stripping punctuation and pronouns, and reducing all words to their respective word stems. After this filtering process, we retain approximately 45,819 unique tokens.

This cleaned descriptions corpus is decomposed into book genres using the previously mentioned LDA model. LDA provides a statistical framework for generating documents based on topics. It is an unsupervised topic model that clusters words into topics (or genres), which are distributions over words, and classifies descriptions as mixtures of these topics/genres. The term *latent* is used because the words are intended to communicate an underlying structure, namely, the subject matter (topic) of the description. The term *Dirichlet* is used because the topic mixture is drawn from a conjugate Dirichlet prior (Thorsrud 2020).

The structure of the LDA model is as follows: the whole corpus is represented by M distinct documents (descriptions) and $N = \sum_{m=1}^{M} N_m$ is the total number of words in all documents. Assuming K latent topics/ genres, each topic is given by a probability vector $\boldsymbol{\phi_k} = (\phi_{k,1}, ..., \phi_{k,N})$ with $\sum_{n=1}^{N} \phi_{k,n} = 1$ indicating the probability that each word shows up in this topic. Further, each document $m \in \{1, ..., M\}$ contains all topics with different probabilities (weights) $\boldsymbol{\theta_m} = (\theta_{m,1}, ..., \theta_{m,K})$ with $\sum_{k=1}^{K} \theta_{m,k} = 1$. Both $\boldsymbol{\phi_k}$ and $\boldsymbol{\theta_m}$ are assumed to have conjugate Dirichlet distributions with hyperparameters (vectors) α and β, respectively.

Given $\boldsymbol{\phi_k}$ and $\boldsymbol{\theta_m}$, a document is generated by drawing for each word a topic $k \in \{1, ..., K\}$ according to the probabilities $\boldsymbol{\theta_m}$ and one word from the selected topic according to its distribution $\boldsymbol{\phi_k}$. This procedure is repeated until the length of the document is reached. To solve the LDA model, we a priori set $\alpha = 50$ and $\beta = 0.01$. The hyperparameter optimization is executed by using Gibbs simulations. Gibbs sampling (also known as alternating conditional sampling) is a specific form of Markov chain Monte Carlo and simulates a high-dimensional distribution by sampling on lower-dimensional subsets of variables where each subset is conditioned on the value of all others (e.g., Steyvers and Griffiths 2007).

The sampling is performed sequentially and proceeds until the sampled values approximate the target distribution. We set the number of sampling iterations equal to 1000. Then, based on different coherence values across the estimated LDA models with varying numbers of genres, we found that 15 topics/genres provide an optimal statistical decomposition of our book description corpus. This number also provides a comprehensible set of topics that can

[27] The base module is gensim by Řehůřek and Sojka (2010), an open-source NLP text analytics tool.

Table 12 OLS estimation approach for different genres generated by LDA or given by Amazon. Dependent variable is the logarithm of the retail price for e-books sold on Amazon

	Dependent variable: log Price				
	(1)	(2)	(3)	(4)	(5)
Agency	−0.204***	−0.210***	−0.210***	−0.210***	−0.202***
	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)
log sales rank	0.056***	0.055***	0.060***	0.059***	0.051***
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
log Kindle Size	0.045***	0.045***	0.050***	0.050***	0.027***
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
log star rating	0.502***	0.492***	0.491***	0.484***	0.436***
	(0.048)	(0.049)	(0.048)	(0.048)	(0.048)
No. expert reviews	0.013***	0.014***	0.015***	0.015***	0.018***
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
log RRP	1.090***	1.094***	1.076***	1.071***	1.080***
	(0.011)	(0.011)	(0.011)	(0.012)	(0.012)
Date Retail	0.011***	0.011***	0.013***	0.012***	0.008***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Bestsellers	0.001***	0.001***	0.001***	0.001***	0.001***
	(0.0004)	(0.0004)	(0.0004)	(0.0004)	(0.0004)
WeekInChart	0.007***	0.007***	0.007***	0.007***	0.007***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Constant	−1.987***	−1.960***	−1.970***	−2.204***	−1.745***
	(0.074)	(0.074)	(0.075)	(0.084)	(0.074)
No. Genre Topics	9	10	12	15	Amazon Genres
Observations	12,001	12,001	12,001	12,001	12,001
Adjusted R^2	0.608	0.608	0.610	0.610	0.615
Robust standard errors in parentheses	$^{*}p$ <0.1; $^{**}p$ <0.05; $^{***}p$ <0.01				

be attributed to officially available genres.[28] A detailed list of all 15 genres is presented in Table 13.

A caveat of the LDA estimation procedure is that it does not assign names or labels to the topics/genres. Thus, labels are subjectively assigned to each genre based on the most important words associated with each topic. In most cases, classifying the genres is conceptually straightforward. Besides, the exact labeling plays no material role in our empirical approach; it is merely a convenient way of referring to the different topics instead of solely using topic numbers.

It is more important that the LDA decomposition yields a meaningful and easily interpretable genre classification of the book descriptions. Our LDA approach achieves this by identifying all important book genres and clearly delineating the topics. This is demonstrated in Figs. 10, 11, and 12, which illustrate three examples of our 15 word clouds. These clouds

[28] For 15 genres, the coherence value exhibits a local optimum when the elbow method is applied (Thorndike 1953). We also considered 9, 10, and 12 genres, but the agency effect remained robust when controlling for a different number of genres (see Table 12).

Table 13 Fifteen different genres identified by our LDA approach

Topic	Genre
0	Philosophy
1	Textbooks
2	Childen and Youth
3	History
4	Guidebook
5	General Fiction
6	Fantasy Fiction
7	Personal development
8	Cultural
9	Self-Guidance
10	Thriller
11	Historic Novel
12	Social Novel
13	Crime Novel
14	Politics

visualize the genre distribution of words by their assigned probabilities through the LDA. We have labeled the topic in Fig. 10 as *Crime Novel*, in Fig. 11 as *Thriller* and the genre in Fig. 12 as *Politics*. In these clouds, a larger word size indicates a higher weight within its respective topic.

The LDA allows assigning each document (book description) from our corpus (data) a probability of belonging to each respective topic. For visualization in Fig. 13, which builds upon Fig. 1, the largest probability value is selected to highlight the distribution of topics across e-book prices and publishers. This distribution of prices by genre demonstrates high comparability between the publishers in our data set because they are not specialized in

Fig. 10 Topic Crime Novel from the LDA. Size is according to weight of word within the topic

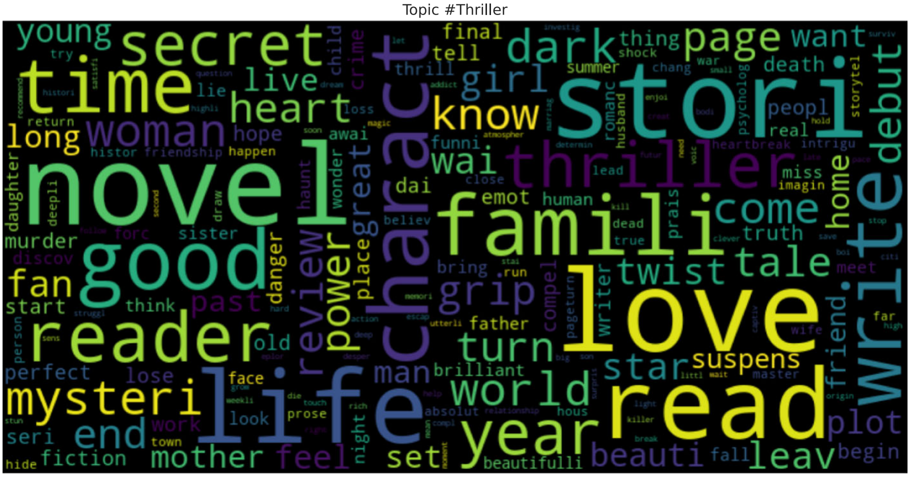

Fig. 11 Topic Thriller from the LDA. Size is according to weight of word within the topic

certain topics, but all publishers sell book titles from different genres. This is an advantage for our empirical approach, as it mitigates potential multicollinearity issues. From an economic perspective, it ensures we can meaningfully compare publishers, avoiding a scenario where specific topics were exclusively from specific publishers (which would preclude comparison).

However, it is evident that individual publishers exhibit distinct primary topics. For instance, Pan Macmillan primarily publishes fiction titles like crime novels, thrillers, or society novels, whereas HarperCollins focuses on family novels and drama genres. However, it is important not to take these topics at face value, as the LDA assigns a probability to each topic.

Fig. 12 Topic Politics from the LDA. Size is according to weight of word within the topic

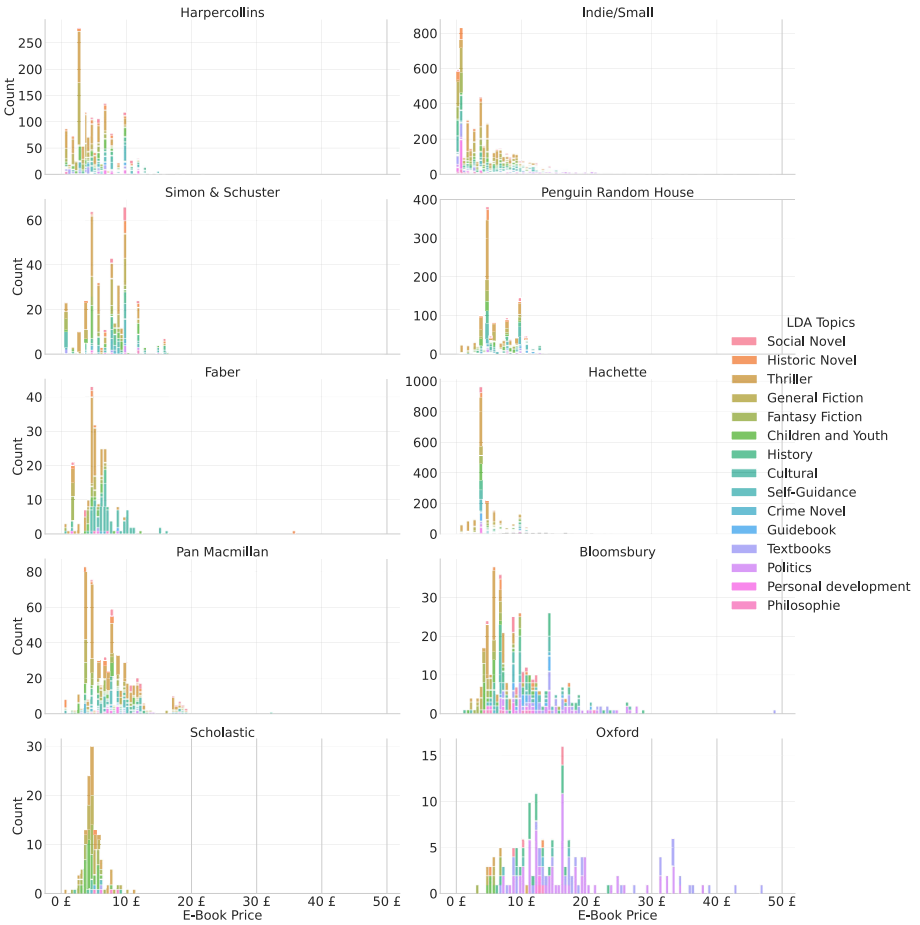

Fig. 13 Prices for e-books grouped by publisher and genre. The ordinate is scaled differently for each subplot

Acknowledgements We would like to thank participants at the MaCCI Annual Conference 2021, EARIE Annual Conference 2021, and CRESSE Conference 2021. In particular, we thank Georg Götz, Daniel Herold, Jan Thomas Schäfer, Jona Stinner, Xiang Hui, Franco Mariuzzo, and Matthew Olczak for helpful comments. The authors alone are responsible for the content.

Author Contributions M.M.G.: Methodology, Software, Validation, Formal Analysis, Investigation, Data Curation, Visualization, Writing – Review & Editing.
P.-A.K.: Conceptualization, Methodology, Validation, Formal analysis, Writing – Original Draft, Supervision.

Funding Open Access funding enabled and organized by Projekt DEAL. No Funding received for this research.

Data Availability Data will be made available on request.

Declarations

Ethical Approval N/A

Informed Consent N/A

Conflict of Interest The authors declare no competing interests.

References

Abadie A, Imbens GW (2006) Large sample properties of matching estimators for average treatment effects. Econometrica 74(1):235–267

Abadie A, Imbens GW (2011) Bias-corrected matching estimators for average treatment effects. Journal of Business & Economic Statistics 29(1):1–11

Abhishek V, Jerath K, Zhang ZJ (2016) Agency selling or reselling? Channel structures in electronic retailing. Manage. Sci. 62(8):2259–2280

Aguiar L, Waldfogel J (2018) Quality predictability and the welfare benefits from new products: evidence from the digitization of recorded music. J. Polit. Econ. 126(2):492–524

Athey S (2018) The impact of machine learning on economics, The Economics of Artificial Intelligence: An Agenda, 507–547. University of Chicago Press

Athey S, Imbens GW (2019) Machine learning methods that economists should know about. Annual Review of Economics 11:685–725

Austin PC (2011) Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. Pharm. Stat. 10(2):150–161

Battocchi K, Dillon E, Hei M, Lewis G, Oka P, Oprescu M, Syrgkanis V (2019) EconML: A Python package for ML-based heterogeneous treatment effects estimation. Version 0.12.0

Brynjolfsson E, Hu Y, Simester D (2011) Goodbye pareto principle, hello long tail: the effect of search costs on the concentration of product sales. Manage. Sci. 57(8):1373–1386

Brynjolfsson E, Hu Y, Smith MD (2003) Consumer surplus in the digital economy: estimating the value of increased product variety at online booksellers. Manage. Sci. 49(11):1580–1596

Brynjolfsson E, Hu Y, Smith MD (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

Chernozhukov V, Chetverikov D, Demirer M, Duflo E, Hansen C, Newey W, Robins J (2018) Double/debiased machine learning for treatment and structural parameters. Economet. J. 21(1):C1–C68. https://doi.org/10.1111/ectj.12097

Condorelli D, Galeotti A, Skreta V (2018) Selling through referrals. Journal of Economics & Management Strategy 27(4):669–685

Davies S, Coles H, Olczak M, Pike C, Wilson C (2004) Benefits from competition: some illustrative UK cases. DTI Economics Papers No. 9

De los Santos B, O'Brien DP, Wildenbeest MR (2021) Agency pricing and bargaining: evidence from the e-book market

De los Santos B, Wildenbeest MR (2017) E-book pricing and vertical restraints. Quant Mark Econ 15(2):85–122

Dearnley J, Feather J (2002) The UK bookselling trade without resale price maintenance an overview of change 1995–2001. Publ. Res. Q. 17(4):16–31

Ennis S, Ivaldi M, Lagos V (2023) Price-parity clauses for hotel room booking: empirical evidence from regulatory change. The Journal of Law and Economics 66(2):309–331

Faber RP, Janssen MC (2019) On the effects of suggested prices in gasoline markets. Scand. J. Econ. 121(2):676–705

Fishwick F (2008) Book prices in the UK since the end of resale price maintenance. Int. J. Econ. Bus. 15(3):359–377

Foros Ø, Kind HJ, Shaffer G (2017) Apple's agency model and the role of most-favored-nation clauses. Rand J. Econ. 48(3):673–703
Gaudin G, White A (2014) On the antitrust economics of the electronic books industry. DICE Discussion Paper 147 [rev.], Heinrich Heine University Düsseldorf, Düsseldorf Institute for Competition Economics (DICE)
Gentzkow M, Kelly B, Taddy M (2019) Text as data. Journal of Economic Literature 57(3):535–74
Gilbert RJ (2015) E-books: a tale of digital disruption. Journal of Economic Perspectives 29(3):165–184
Greifer N (2024) cobalt: covariate balance tables and plots. R package version 4(5):5
Hunold M (2017) Best price clauses: what policy as regards online platforms? Journal of European Competition Law & Practice 8(2):119–125
Hunold M, Kesler R, Laitenberger U, Schlütter F (2018) Evaluation of best price clauses in online hotel bookings. Int. J. Ind. Organ. 61:542–571
Hunold M, Muthers J (2017) Resale price maintenance and manufacturer competition for retail services. Rand J. Econ. 48(1):3–23
Ippolito PM (1991) Resale price maintenance: empirical evidence from litigation. The Journal of Law and Economics 34(2, Part 1): 263–294
Johnson JP (2017) The agency model and MFN clauses. Rev. Econ. Stud. 84(3):1151–1185
Johnson JP (2020) The agency and wholesale models in electronic content markets. Int. J. Ind. Organ. 69:102581
Kleiber C, Zeileis A (2008) Applied econometrics with R. Springer Science & Business Media, Berlin
Knaus MC (2021) A double machine learning approach to estimate the effects of musical practice on student's skills. J. R. Stat. Soc. A. Stat. Soc. 184(1):282–300
Larsen VH, Thorsrud LA (2019) The value of news for economic developments. Journal of Econometrics 210(1):203–218
Lee BK, Lessler J, Stuart EA (2010) Improving propensity score weighting using machine learning. Stat. Med. 29(3):337–346
Lenz D, Winker P (2020) Measuring the diffusion of innovations with paragraph vector topic models. PLoS ONE 15(1):1–18
Lu L (2017) A comparison of the wholesale model and the agency model in differentiated markets. Rev. Ind. Organ. 51:151–172
MacKay A, Smith DA (2017) Challenges for Empirical Research on RPM. Rev. Ind. Organ. 50(2):209–220
Mantovani A, Piga CA, Reggiani C (2021) Online platform price parity clauses: evidence from the EU Booking.com case. European Economic Review 131:103625
Mathewson GF, Winter R (1984) An economic theory of vertical restraints. Rand J. Econ. 15(1):27–38
Motta M (2004) Competition policy: theory and practice. Cambridge University Press
Nelson P (1970) Information and consumer behavior. J. Polit. Econ. 78(2):311–329
Perry MK, Porter RH (1986) Resale price maintenance and exclusive territories in the presence [of] retail service externalities. State University of New York at Stony Brook, Department of Economics
Poort J, van Eijk N (2017) Digital fixation: the law and economics of a fixed e-book price. International Journal of Cultural Policy 23(4):464–481
Řehůřek R, Sojka P (2010) Software framework for topic modelling with large corpora. In: Proceedings of the LREC 2010 workshop on new challenges for NLP frameworks, Valletta, Malta, pp. 45–50. ELRA
Reimers I, Waldfogel J (2021) Digitization and pre-purchase information: the causal and welfare impacts of reviews and crowd ratings. American Economic Review 111(6):1944–71
Rey P, Stiglitz J (1988) Vertical restraints and producers' competition. Technical report, National Bureau of Economic Research
Rey P, Stiglitz J (1995) The role of exclusive territories in producers' competition. Rand J. Econ. 26(3):431–451
Rey P, Vergé T (2010) Resale price maintenance and interlocking relationships. J. Ind. Econ. 58(4):928–961
Robinson PM (1988) Root-N-consistent semiparametric regression. Econometrica: Journal of the Econometric Society 56(4):931–954
Rosen S (1974) Hedonic prices and implicit markets: product differentiation in pure competition. J. Polit. Econ. 82(1):34–55
Rosenbaum PR, Rubin DB (1983) The central role of the propensity score in observational studies for causal effects. Biometrika 70(1):41–55
Rubin DB (1977) Assignment to treatment group on the basis of a covariate. J. Educ. Stat. 2(1):1–26
Rubin DB (1978) Bayesian inference for causal effects: the role of randomization. Ann. Stat. 6(1):34–58
Sekhon JS (2011) Multivariate and propensity score matching software with automated balance optimization: the matching package for R. J. Stat. Softw. 42(7):1–52
Spengler JJ (1950) Vertical integration and antitrust policy. J. Polit. Econ. 58(4):347–352

Steyvers M, Griffiths T (2007) Probabilistic topic models. Handbook of Latent Semantic Analysis 427(7):424–440
Thorndike RL (1953) Who belongs in the family? Psychometrika 18(4):267–276
Thorsrud LA (2020) Words are the new numbers: a newsy coincident index of the business cycle. Journal of Business & Economic Statistics 38(2):393–409
Tirole J (1988) The Theory of Industrial Organization. MIT press, Cambridge
Varian HR (2014) Big data: new tricks for econometrics. Journal of Economic Perspectives 28(2):3–28
Wang X, Yucesoy B, Varol O, Eliassi-Rad T, Barabási AL (2019) Success in books: predicting book sales before publication. EPJ Data Science 8(1):31
Westreich D, Lessler J, Funk MJ (2010) Propensity score estimation: neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression. J. Clin. Epidemiol. 63(8):826–833
White H (1980) A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. Econometrica 48(4):817–838
Winter RA (1993) Vertical control and price versus nonprice competition. Q. J. Econ. 108(1):61–76
Wooldridge JM (2013) Introductory econometrics: a modern approach, 5th edn. South-Western Cengage Learning, Mason, OH