Have We Misunderstood Copyright's Consequences?

Stan J. Liebowitz

liebowit@utdallas.edu

School of Management

University of Texas at Dallas

This paper uses an unusually rich 21st century data set to compare two sets of vintage bestsellers from the early 20th century that, by a circuitous path of copyright law alterations, came to have different copyright treatments. Surprisingly, copyrighted works are found to sell almost four times as many copies as public domain works, not just on average but throughout the sales distribution, contradicting the expectation that copyright would restrict sales. These greater sales occur despite a price premium that we find for copyrighted works, which on average is of a size similar to typical royalty payments, although the premium is considerably higher for better selling editions. We also find, contrary to previous claims, that copyrighted titles are more likely to be sold than are titles in the public domain. Copyright, therefore, is more likely to be socially beneficial than previously thought, retroactive copyright extensions are more likely to be socially advantageous, and indefinitely renewable copyright is more likely to be an optimal policy.

Electronic copy available at: https://ssrn.com/abstract=4547458

Copyright provides ownership over creative works. It is the system that market economies generally adopt with the putative goal of providing creators with an incentive to generate new artistic works. Without copyright, unfettered competition within the market for copies of a particular title would be expected to drive the economic profit from book publishing to zero, leaving no revenue with which to pay the author. Copyright is typically believed to provide a balance, often thought by critics to be lopsided, between the positive inducement of producing new creative works and the negative restriction on the consumption of protected works. Although copyright does not restrict competition between titles, it has nevertheless been expected to decrease sales while increasing price (and profit) for works to which it is assigned, as would be expected from a canonical monopoly. This welfare balance, lopsided or not, is often referred to as the "access/incentive" tradeoff.

My goal here is to examine copyright's impact on the sales, availability, and price of books, ignoring the effects of copyright on the creation of works. I take advantage of a natural experiment where, due to a quirk in the serial alterations of America's copyright law, titles written prior to 1923 lost their copyrights while titles written after 1922 were able to retain theirs. Due to the difficulty of acquiring data on the sales of copyrighted products, there have been very few prior comparisons of these groups of titles. One of the novelties of my analysis is the use of a data set (Nielsen BookScan, now called NPD BookScan) that contains information on the sales of individual titles, whereas previous analyses, few though they may be, mostly used Bowker's Books in Print (BiP) which contains no sales data, or, if sales data were used, they tended to come from a single retailer.

Using BookScan sales data reveals the surprising finding that copyrighted (CR) titles sell many more copies than public domain (PD) titles. This noteworthy sales advantage exists throughout the sales distribution, including both high and low selling titles. My analysis also finds that CR titles are sold at higher prices than PD titles, with the average price difference is in the vicinity of royalty rates that publishers pay authors, although the more successful titles have considerably higher copyright premiums. Additionally, my analysis finds that copyrighted titles are slightly more likely to be sold than are public domain titles, in contrast to comparisons based on BiP measures of "in-print" availability which turns out to be an inadequate predictor of market activity.

The greater sales exhibited by copyrighted works suggest that copyright is poorly explained by the monopoly model. This is not surprising since copyright merely provides ownership, which should not be, but often is, conflated with monopoly in this literature.[1] If copyright increases sales, it is possible, perhaps likely, that copyright provides social benefits other than just

[1] Economists and others (e.g., Boldrin and Levine, 2008) commonly refer to copyright as providing a governmentally authorized monopoly to creators, although Kitch (2000) argues that copyright merely provides ownership, not monopoly, and Liebowitz (2016b) points out that ownership can always be viewed as a tautological monopoly though usually one with no monopoly power.

Electronic copy available at: https://ssrn.com/abstract=4547458

providing authors a greater financial incentive to produce new works. Instead of balancing the harm from monopoly against the benefit of new creations, both production and consumption would be enhanced by copyright, at least with respect to purchase of copies of the work by the public.[2]

One explanation for why sales quantities might be enhanced by copyright is simply that copyright provides the ownership rights that allow a publisher to internalize all the returns from its marketing investments (Landes and Posner [2003], Adilov and Waldman [2013]).[3] Without copyright, a publisher trying to foster market demand for a book title would likely have to share any returns with free-riding publishers selling the same title, reducing the incentives to make such marketing investments. The public domain's disadvantage in this case would simply be another example of the tragedy of the commons, where the lack of property rights for public domain works fails to provide efficient incentives for sales-promoting investment, leading to inefficiency.

When performing this analysis, a specific nomenclature is required to keep track of the units of observation. Individual creations, such as Steinbeck's *Of Mice and Men,* are referred to as titles. There can be multiple variations of a single title, such as hardcover or paperback versions, or versions from different publishers if the title is not copyrighted. The various versions of a title are referred to as "editions" and there can be dozens of contemporaneous editions of a single title, particularly for popular public domain titles.

I. Copyright and the Public Domain

Copyright owners are granted the exclusive right to make reproductions of their work (title) and that right has been extended to various forms of 'reproduction' including public performance, derivative works, electronic transmissions to the public and so forth, but it does not provide any protection from independently created competing works.

Book publishing is the oldest industry relying on copyright. After Gutenberg invented the printing press in the mid-1400s, various European governments granted printing monopolies, such as the monopoly given to the Stationer's Company (essentially a guild of printers) in England, chartered in 1557. The monopoly on printing in England was ended in 1709 by the first modern copyright

[2] This analysis ignores copyright's impact on the creators of follow-on works trying to obtain the permission of the copyright owner. Thus, copyright could reduce the production of new follow-on works, and since the elements of the original work that might be "borrowed" are nonrivalrous, any exclusion would be inefficient. Because it is unlikely that competing creators of follow-on works could engage in arbitrage, the copyright owner should be able to engage in price discrimination and price discrimination can cure imperfections from market production of nonrivalrous goods. Nevertheless, this is a complex topic best handled at another time and I am willing to stipulate that my analysis here is limited to the primary market for copies of the work sold to consumers.

[3] The term "marketing" here is meant to convey prosaic activities such as getting books into bookstores, on educational reading lists, getting mentions in popular articles, direct advertising, and so forth.

Electronic copy available at: https://ssrn.com/abstract=4547458

law, the Statute of Anne, which provided a copyright for titles, with a term of 14 years renewable for an equal second term.

The original American copyright law of 1790 was modelled on the Statute of Anne and provided 14 years of protection for American authors, followed by another round of 14 years if the copyright owner renewed the protection. In 1831, the length of the first term was increased from 14 years to 28 years. In 1909 both terms were set to 28 years. In 1964 an additional 19 years was added to that second term for all then-current and future copyrighted (CR) works. Thus, a title first published in 1922 (and renewed in 1950) would have been expected to enter the public domain 75 years later (28+28+19) in 1997, and similarly, a title published in in 1923 would have been expected to enter the public domain in 1998.[4] In 1998, however, Congress added 20 years protection for works that were then currently under copyright. Hence, a title published in 1923 (or later) and then renewed, would have continued to be protected until 2018 whereas a title published in 1922 (or before) would have lost protection in 1997. *Thus, for our data, which examines the 2004-2016 sales of titles originally published from 1895 until 1950, there are no titles with a change in copyright status.*

This quite arbitrary 1922/23 dividing line between titles with and without copyright provides a natural experiment that we will use to identify the consequences of copyright protection on the price and sales of books. The logic is quite simple. Titles that are still being published a century or so after being written have all proven the ability to retain the interest of generations of readers. Whether a title was written in 1922 or 1923, however, is nothing more than a random accident that has nothing to do with the titles themselves. Those titles published in 1923 or later have, in years prior to 2018, received the "treatment" of copyright, whereas those written prior to that date were not so treated.

II. Prior Literature

Until the last decade and a half, there were virtually no empirical studies examining the impact of copyright on book prices, sales, or availability. Two studies undertaken at about the same time, Heald (2008) and Liebowitz (2009), looked at how the 21st century prices of titles that were first published in the period surrounding the 1923 copyright cutoff, compared to one another. The evidence in those studies is based mainly on the BiP data set which provides information (e.g., edition year, price, pages, format, publisher) about editions of book titles. If an edition is listed as "in-print" it is treated as available for sale to the public.

Heald (2008) chose old titles from three samples. The main sample was based on 20 years of bestsellers around the 1923 cutoff, but he supplemented the sample with 40 titles written during that period which were not necessarily bestsellers at the time but were popular in recent decades. He used data taken from the 2006 online edition of BiP (and offline BiP editions from

[4] Although there was an important change to copyright law in 1976 that altered the term of copyright for new works, it did not affect the copyright duration of the old works discussed here.

Electronic copy available at: https://ssrn.com/abstract=4547458

previous years and decades) to compare public domain and copyrighted titles. He was most interested in whether PD titles were more likely or less likely to be available to consumers than CR titles, but he also tried to determine whether there was a price difference between PD and CR titles. When a title had multiple editions he chose the lowest price among the various editions, and he compared the average of these "lowest price" variants for his sample of CR and PD titles. For his complete sample of 287 titles still in-print, he found no copyright price premium. By way of contrast, however, his supplemental sample of 40 recently popular durable titles (that need not have been bestsellers when published) he found that copyrighted titles had considerably higher prices, with the copyright premium ranging from 41% to 81%. Because he used the lowest priced edition as the price for a title instead of some sort of average price of editions, and because he did not control for number of pages or the format (e.g., hardcover), it is not clear how useful his results are. A perhaps more compelling but less general result was his comparison of prices per page for Penguin Classic paperback titles, which implicitly controls for format, pages, and publisher. He found that the copyrighted editions were 56% more expensive per page.

Liebowitz (2009), in a preliminary version of this work, focused on the price differential that copyright provided. Using the hardcover 2004-5 BiP, the 1923 copyright cutoff, and including titles from 1895-1940, he used a regression analysis with the list price as the dependent variable and independent variables such as number of pages, format type, genre, publisher type, copyright status and a sales proxy (Amazon ranking) that we now know to be unreliable (Liebowitz and Zentner [2023]). Like Heald, he found that copyright had no effect on price for the full sample. He also found that limiting results to major publishers did not alter his results.

A decade later Reimers (2019) used a similar methodology (categorized as a regression discontinuity), comparing the prices of editions of former best-sellers published between 1910 and 1936 and using the 1923 copyright demarcation. She examined copyright's impact on the Amazon price for these editions, controlling for several factors. Using Amazon prices seems to be an unusual choice since the publisher doesn't set the price that retailers charge and the question of interest is the price set by the publisher.[5] Of greater consequence was her decision to include both new and used books in her analysis, as we will see below. Nevertheless, she concludes that CR editions are priced 27% higher overall. Additionally, she finds no support for the suggestion that CR might increase a publishers' post-creation investment in a title, and that PD titles had a much larger number of editions. Finally, and the main focus of her analysis, was her conclusion that a retroactive copyright extension, such as the 1998 Copyright Act, was generally harmful to social welfare if the title-production incentive impact of copyright were ignored.

[5] The determination of retail prices came to a head in United States v. Apple Inc., 952 F. Supp. 2d 638 (S.D.N.Y. 2013) whereby leading publishers had tried to switch to an "agency model" allowing the publishers to determine the price to customers with companies such as Apple taking a percentage of revenue, as is done in app stores. If Amazon did not switch to this model, the publishers threatened to withhold books. The Court found this behavior to be an antitrust violation. Similarly, publishers do not control the price that bookstores charge to customers.

Electronic copy available at: https://ssrn.com/abstract=4547458

Finally, using historical data, Li, MacGarvie, and Moser (2018) examine how an 1814 change in English copyright law altered the relative price of editions depending on whether authors died or survived during the initial copyright period, with the editions from surviving authors having a smaller increase in CR duration from the changed law. This is not a comparison of the price of titles with or without CR, but a comparison of prices based on the remaining copyright *length* of the title. Prices are affected if publishers price discriminate during the life of a copyrighted title by starting with a high initial price and lowering it as the remaining time under CR decreases. Li, MacGarvie, and Moser conclude that the greater extension of copyright for works of dead authors raised the price of those books by about 43% relative to the works with a shorter increased CR length. It is not clear whether these values are at all comparable with copyright's impact in modern markets since the market organization was vastly different in early nineteenth century England with, for example, book customers being mainly for-profit lending libraries, not individuals.

III. The book publishing industry

Book publishing can be decomposed into several major categories, with "trade" books (books purchased by individuals) making up the largest share. Trade books are further classified as adult fiction, adult nonfiction, and juvenile. Both BookScan and Books in Print (BiP) provide information on new physical titles (versus used books) in each of these three categories. Although there is a used book market, copyright owners and their publishers sell new books.

There are many book publishers, with over 2200 listed in the 2004 Nielsen BookScan data. Imprints, which are smaller publishing units within major publishers, number almost 7500.[6] Nevertheless, the largest publishers tend to dominate the market. For the 2004-2016 period, the top 4 publishers generated more than 75% of revenue in the market for physical works of fiction and the top 20 publishers generated more than 97% of the revenue. Comparable values for the largest imprints are 15% and 44%. Our definition of "major publishers" is based on the 9 largest publishers in terms of fiction sales,[7] although it is sometimes unclear what the definition of major publishers is in the literature.[8] As we discuss below, sales per edition for major publishers is much larger than average edition sales for minor publishers and this difference is even stronger for our sample of vintage titles.

[6] These numbers overstate the number of independent publishers (and understate concentration) because oftentimes two publishers under the same ownership are listed as separate in the BookScan data although I tried to merge some leading publishers known to be under single ownership (e.g., Random House and Penguin) when calculating the market share values.

[7] We determine "major" publishers as the top selling fiction publishers in 2004 that are not specialized in self-publishing. These include, in order of sales, Random House/Penguin Group, Simon & Schuster, Hachette Book Group, Harpercollins, Macmillan, Harlequin Books, Houghton Mifflin Co, Kensington Publishing, and WW Norton, each of which had 1% or more of the market in terms of revenue and together represent 72% of market sales.

[8] For example, I could not find the definition used in Heald or Reimers.

Electronic copy available at: https://ssrn.com/abstract=4547458

The business model of printing books has changed in some important ways over the last few decades. Offset printing, also known as lithography, has been the primary form of book printing during the last century, with the disadvantage of relatively high setup costs but the advantage of low variable costs. Print on demand (POD) digital technology is a newer printing method with lower fixed costs but higher variable costs (and essentially zero inventory costs because inventories are not needed). Book editions with low expected demand often use POD whereas editions with high expected demand usually use offset. Although another "printing" change in the 21st century has been the introduction of eBooks, which are not included in BookScan data, such books did not have a measurable market presence in 2004 (or 2008), the focal year of this study.

POD has allowed for new business models where small publishers can have vast libraries of titles with no inventory and virtually no fixed printing costs since the book is not printed until a copy is ordered. Some of the on-demand publishers, such as *IndyPublish*, *Books on Demand*, *BiblioBazaar*, and *Kessinger*, often list many more editions and ISBNs than do more established publishers with much larger sales.[9]

For example, Publishers Weekly in 2010 reported:

> …many in the industry were stunned to see an unfamiliar company name, BiblioBazaar, leading a surging new segment of "non-traditional" publishing stats with a whopping 272,930 titles produced in 2009--almost as many titles [as] the entire "traditional" publishing business cranked out last year. Could it be? Could one little-known company really produce so much volume?[10]

> …"If by 'produce' you mean create a cover file that will print at multiple POD vendors, a book block that will print at multiple POD vendors, and metadata to sell that book in global sales channels, then yes, we did produce that many titles," said Mitchell Davis, president of BiblioLife, parent company of BiblioBazaar.[11]

These small "nontraditional" publishers pose a danger to researchers examining titles in-print. I use the term "danger" because many editions from these publishers, perhaps most, do not appear to sell any copies in a year or even a decade, making them more like phantom editions

[9] Most published editions of titles in the US receive an ISBN (International Standard Book Number) number purchased from Bowker Identifier Services. An ISBN represents a single format or a single edition of a title. Some self-published authors choose not to use an ISBN number.

[10] Publishers Weekly also reported that the second and third highest production of titles was from Books LLC and Kessinger Publishing LLC which produced 224,460 and 190,175 titles respectively. See https://www.publishersweekly.com/pw/by-topic/industry-news/publishing-and-marketing/article/42826-self-published-titles-topped-764-000-in-2009-as-traditional-output-dipped.html

[11] See https://www.publishersweekly.com/pw/by-topic/industry-news/publisher-news/article/42850-bibliobazaar-how-a-company-produces-272-930-books-a-year.html

Electronic copy available at: https://ssrn.com/abstract=4547458

than real editions. This is not merely a hypothetical concern: 480 out of 774 BiP listed editions in my soon to be described vintage bestseller data, failed to sell any copies in 2004, and 311 of these editions failed to sell any units in the full 13 years of my data. A large portion of the zero-selling editions in my sample comes from the 310 Kessinger Publishing Company editions, all of which had zero sales in 2004 (and 163 of which were zero-sellers over the full 13-year interval of my data). The data used by Reimers are strongly influenced by these types of editions.[12]

Treating these non-selling editions as if they were being sold could have powerful misleading effects on conclusions that researchers might draw. Small publishers such as these might have unusual pricing strategies, as Ellison and Ellison (2018) discuss for sellers of used books. For example, in my vintage bestseller data, the publisher Kessinger is listed in BiP as providing 117 different paperback editions at an abnormally low price of $1.99 (see Table 1 below for typical paperback prices)[13] and Reimers' data include 68 such editions. Because Heald (2008) uses the lowest price edition (whether sold or not) as the representative price for a title, if any of these Kessinger editions were in in his data, his methodology would conclude that the selling price for these titles was much lower than the prices consumers actually paid when they purchased editions of these titles.[14]

Another publisher specializing in old titles is Reprint Services Corporation. Unlike the Kessinger editions previously mentioned, however, their titles tend to have unusually high prices, averaging $96, and most of their titles in my sample (23 out of 26) do not sell any units in 2004 and 17 sell zero units in the full 13 years. As was the case with Kessinger editions, an analysis that treats these editions as equivalent to editions that actually are sold in markets is likely to distort results.

IV. The Data

My sample of former bestsellers consists of the top 10 fiction best-sellers from Bowker's *Publishers Weekly* yearly bestseller lists in each year from 1895-1950,[15] for a seeming total of 560 possible titles. The actual number of titles differs slightly from this value, however, because

[12] For example, among the 2,862 active hardcover and paperback editions in Reimers's data, 1043 were from BiblioBazaar and 590 were from Kessinger. Furthermore, only 493 of these 1633 editions (~30%) sold any copies over the 13-year period of my BookScan data.

[13] None of these 117 editions sell any copies during the entire 13-year period of the BookScan data even though some of these editions covered well known titles with high yearly sales.

[14] This is consistent with the substantial increase in CR vs PD price differential he found after excluding minor publishers.

[15] Titles published between 1895 and 1909 originally had a copyright term of 28 years plus an additional 14 if renewed, as opposed to the 28+28 terms available during and after 1910. Those pre-1910 titles might, in theory, have been expected to be slightly stronger works on the margin since to cover a fixed opportunity cost of writing, the expected payout per year on a marginal title would need to be higher with fewer years of payout. Given my focus on bestsellers, however, I would not expect this minor difference to play a role in the analysis. If it did, however, it would advantage sales of works in the public domain.

Electronic copy available at: https://ssrn.com/abstract=4547458

individual titles are often listed in more than one year's best-sellers list, and sometimes a yearly tie in ranking leads to 11 titles being listed in a year. I will hereafter apply the term "vintage" to these former bestsellers.

Once this set of vintage best-selling titles was created, 21st century information was collected for each edition of each title using the 2004-5 hardcopy version of Bowker's "Books in Print" database which does not contain sales data. Because I only had BiP data for 2004, that became the focal year for the analysis. Many of the century old best-sellers were no longer in-print in the early 21st century, according to BiP. For those that were in-print, my RAs examined titles published after 1922 and determined whether their copyright was renewed 28 years later.[16] All post-1922 titles that were renewed were deemed copyrighted, all other titles in the sample were deemed to be in the public domain.

I also had access to Nielsen (now NPD) BookScan data, covering the years 2004-2016, for physical copies of works of fiction that were sold by both online and offline retailers.[17] The BookScan data set contains information for all physical editions sold in BookScan's large panel of retailers, thought to represent 85% of the market, which, given its coverage of retailers carrying large inventories, is likely to include just about every edition (but not every unit) sold by retailers in the U.S. BookScan data includes, in addition to sales numbers, edition information similar to what is found in BiP.[18]

BookScan is considered the gold standard for data on physical book sales. The quality of the BookScan data is attested to by the fact that BookScan's subscribers are mainly book publishers wishing to keep track of how their titles are doing in the retail market. Publishers are willing to pay for this information since they do not know the actual sales of their titles until retailers finish returning the copies that do not sell, often many months after the initial delivery. Further evidence of the regard in which the BookScan data set is held is the fact that Amazon, wishing to provide authors with detailed information about their book sales, gives them access to BookScan

[16] Examining renewals required checking each post-1922 title against the *Catalog of Copyright Entries,* and an online database of that information can be found at https://exhibits.stanford.edu/copyrightrenewals. They found that 3 of the 180 post-1922 titles had lost their copyright in this manner. These were *Black Oxen*, *Mistress Wilding*, and *The Plastic Age*, published in 1923, 1924, and 1924, respectively.

[17] BookScan derives its data from transaction information (e.g., checkout scanners) reported by many retailers, including online retailers such as Amazon. As not every retailer is included, it is likely that the BookScan sales data will undercount the sales of each title/edition.

[18] I found BookScan data to sometimes contain editions of old best-sellers that were not listed (or not found by my RAs) in my hardcover BiP (BookScan also listed editions not reported by Reimers in her examination of the electronic BiP). I used BookScan data when BookScan and BiP overlapped in their coverage of variables as I deemed electronic data from BookScan likely to be more reliable than hand copied (very small print) numbers from the hardcopy BiP. Although I had BookScan page number data for the small number of vintage bestsellers, I had to use page data from the WorldCat library database when examining the entire fiction sector.

Electronic copy available at: https://ssrn.com/abstract=4547458

sales data.[19] Nonetheless, some important channels, such as book clubs, do not have their sales counted by BookScan, and BookScan is limited to enumerating sales of physical copies. Also, due to limitations in the BookScan data available to me, some of the slowest selling editions are not included in the later years of my data, beginning in 2010. By 2016, almost one third of editions (selling one or two units per year each) are missing.[20] Although these missing editions have a minimal impact on overall sales, the large number of missing editions imply that any estimates that are sensitive to the number of editions are suspect in the latter years of the data.

The vintage best-sellers are currently published using the same printing technologies as recently written titles. In addition, editions of these vintage titles are often published by the same publishers who mainly publish recent titles. Thus, in addition to our sample of vintage titles, we can use information from the very large number of recently written titles to better gauge the book characteristics that determine the price of books currently sold.

The most important influence on price is usually the type of book format. The distinction between hardcover and paperback formats is well-known but there are three major formats used in this market with the third category being mass market paperback. Mass market paperbacks are cheap small books that are made for portability and mainly sold through different outlets than the other two categories.[21] While other obscure types of formats exist in this market, they represent a very small portion of the market and are removed from the analysis.[22]

Since the price of books is an important element in our analysis, it is worth noting that there are some editions that have unusual prices.[23] For example, the best-selling author John Grisham has leather-bound and autographed editions of many of his novels, and they sell for the very high price of $250. I label all unusually high-priced editions as "collectibles" (defined as having a price greater than $65)[24] and remove them from the analysis. There were only 902 collectable editions

[19] This practice began in 2011 and continues to this day. See https://latimesblogs.latimes.com/jacketcopy/2010/12/amazon-gives-nielsen-bookscan-to-authors.html

[20] The BookScan data to which I had access has a limitation of 250,000 editions per year for the major book categories such as Adult Fiction. This limitation began to bind in 2010 when editions with sales of 1 unit began to be missing from the data. By 2016, editions selling 1 unit were completely missing and editions selling two units were largely missing as well. Liebowitz and Zentner (2023) estimate that these missing editions represented about one third of all editions sold in 2016, or about 120,000 editions.

[21] MM paperbacks are largely sold in non-bookstore outlets such as drugstores, supermarkets and airports, play a diminished role over time, and play almost no role in juvenile or nonfiction markets.

[22] Other format/binding categories include "library," "boxed set," and "board books," among others. I merge library with hardcover titles and remove editions in the other categories.

[23] The highest prices for editions sold in 2004 (excluding boxed sets) are $1845, $900, and $750. It is possible that some of these values are data errors.

[24] The 99th percentile price was $51 for the entire market. Vintage editions, when ranked by price, had a large jump from $56 to $67. Given this, the $65 value seemed like a reasonable cutoff. None of the main results in the paper are affected by these choices since most of the high-priced vintage titles had no sales and titles with prices greater than $65 never surpassed yearly sales of 10 units.

Electronic copy available at: https://ssrn.com/abstract=4547458

out of more than 125,000 editions sold in 2004. Similarly, I remove editions with the price of zero which I take to indicate an error in the data.[25]

Table 1: 2004 Format Shares, Revenues and Prices for New and for Vintage Editions Sold

	All editions with +sales & Price<65 >0			Vintage editions with sales>0 & Price<65		
	Share of Editions	Share of Revenue	Avg Price	Share of Editions	Share of Revenue	Avg Price
MM paper	23.4%	29.0%	$6.52	12.0%	22.8%	$6.24
Trade paper	53.8%	35.2%	$16.13	56.2%	73.3%	$17.51
Hardcover	22.8%	35.8%	$25.34	31.9%	3.9%	$26.14
Total	126,202	$2,423,907,707		552	$15,056,851	
Revenue is calculated as the product of unit sales and list price.						

Table 1 compares editions of vintage titles to editions of mainstream fiction titles. The left-side portion of the table, representing the entire adult fiction market in 2004, shows for each format type the share of market by edition and revenue, along with average prices. The right-side portion of the table shows similar statistics for the much smaller number of vintage bestsellers. This comparison reveals that revenues are fairly evenly split between formats for the market as a whole, whereas the vintage titles generate almost no revenue from hardcover editions (in spite of the large number of available hardcover vintage titles and their higher prices) and almost three quarters from trade paperbacks.[26] With regards to price, the vintage editions have prices that are quite similar to the mainstream editions being sold: hardcover editions are about $10 more expensive than trade paperbacks, which in turn are about $10 more than mass market paperbacks which have prices of about $6.

Table 2 compares characteristics of copyrighted and public domain vintage editions. The characteristics for CR and PD editions differ in some dimensions. Copyrighted editions have prices that are 16% higher than the prices of PD editions ($20.63/$17.86).[27] But CR editions also have more pages, and are more likely to be published in hardcover, both of which tend to raise prices. On the other side of the ledger, CR editions are more likely to be published by major publishers, and be older (not the title, but a particular edition), each of which tends to lower price.

[25] Because BookScan data cover physical books that are costly to produce and are sold in (online and offline) retail stores, it is very unlikely that the price would be zero since that would impose potentially large losses on the publisher. There were only 322 instances where 2004 fiction titles had a price of zero (out of 141,042).

[26] Revenue is calculated as the product of list price and quantity sold. Although it is difficult to find data about how many books sell at list price, it is generally agreed that Amazon is a leading book discounter. Using Reimers' data on Amazon book prices in 2012 and comparing them to the 2012 BookScan list prices for a set of 160 editions of vintage bestsellers, I discovered that 52% of editions sold at list price, the average discount was 10%, and that discounts as large as 25%-50% existed for about 20% of editions. Better selling editions tended to have larger discounts. I interpret this to mean that a larger majority of copies are likely to sell at list price since other bookstores are likely to discount less than Amazon.

[27] The copyright premium is 22% for MM Paper, 16% for Trade Paper and 7% for Hardcover.

Electronic copy available at: https://ssrn.com/abstract=4547458

Table 2: Characteristics of 2004 Editions of Vintage Titles Sold, P<65

		mean	t-test Diff	Obs.
Price	PD	$17.86	3.2	342
	CR	$20.63		210
# of Pages	PD	328.4	4.4	338
	CR	396.1		198
Major Publisher	PD	22.8%	3.2	342
	CR	35.7%		210
Edition Year	PD	1997.5	4.6	338
	CR	1992.4		209
Paperback	PD	59.4%	1.9	342
	CR	51.0%		210
MM Paperback	PD	11.7%	0.2	342
	CR	12.4%		210
Hardcover	PD	28.9%	1.9	342
	CR	36.7%		210

V. The Identification Strategy

The broad outlines of the methodology have already been given. Ideally, we would like to have two similar and randomly drawn sets of titles, providing copyright protection to one group and denying copyright protection to the other. We would want to limit variations in the physical quality of the books, so we might require that book formats in the experiment be standardized and that book publishers use the same set of printers. We might insist that all books in the experiment be of a similar number of pages, a similar quality of binding, paper, ink, and the use of illustrations. We could then see how the two groups of titles differed in terms of price, sales, and availability.

But even under terms of this ideal experiment, there might be possible problems. First, titles might have different prices even if they were physically identical. Different titles appeal to different types of readers, implying that consumers' reservation prices could well differ across different titles. Clerides (2002), however, has found that for a publisher who gave him access to financial records, book prices were affected mainly by cost shifters, not demand shifters, so demand variation may not be an important factor in book pricing.

Our actual identification strategy tries to follow this basic logic. Authors writing books in the decades surrounding 1923 are likely to have thought that copyright would have long expired on their creations by the early portion of the 21st century. There is no reason, therefore, to think that the nature of their creations should differ by whether they wrote their books before or after the 1923 cutoff. The 21st century copyright assignment should be unrelated to author efforts and

Electronic copy available at: https://ssrn.com/abstract=4547458

product, and this provides the basis for identifying the impact of copyright on price, sales, and availability.

We also control for book characteristics that might influence prices and sales. By using only works of fiction, we reduce the variability in reading audiences and book characteristics compared to using all categories of books (e.g., textbooks). We can control for some of the physical characteristics of books since they are reported in our data, such as the number of pages, the type of format, or the use of illustrations. Other characteristics are more difficult to control for, such as the quality of the paper, ink, and binding. For these latter characteristics we use fixed effects for the imprint producing the title, in the belief or hope that imprints tend to publish books of a similar physical quality. Supporting this belief that imprints might tend to focus on particular production qualities is evidence that imprints do tend to specialize in editions of particular format-types.[28]

Nevertheless, there are potential problems with this identification strategy as regards the sales of titles. For one thing, the definition of "book" has changed in this century. Although books have traditionally been printed exclusively on paper, in the last few decades digital books have entered the market. This creates two possible issues for our analysis. First, the BookScan data do not include digital editions, so we are ignorant about part of the market, although this will not affect our results in the focal year of 2004 since digital books were of trivial importance. Second, as suggested by Reimers, because there is essentially no cost in producing additional digital copies, and anyone can publish PD titles, digital PD titles published by nonprofits can be priced much lower (at zero) when competing with physical PD titles. This would be expected to reduce the sales of physical PD titles to the extent that nonprofit digital titles are substitutes for physical copies of the same titles.

Another possible problem that to my knowledge has not previously noted is the fact that copyrighted vintage titles in our sample are not randomly assigned with regard to the age of the title but in fact are newer than the public domain vintage titles. Characteristics of bestsellers might change over time in terms of English usage, morals, character stereotypes and so forth, such that the age of a title (its "recency"), which is related to its copyright status, may tend to affect its current sales or availability in the market.

These issues are addressed as we go through the empirics.

[28] If a Herfindahl index is created for the market shares of the three format categories (trade paper, mm paper and hardcover) for the entire market, as found in the left-hand side of Table 1, the index value is 3367. Creating a Herfindahl index for the format shares for individual imprints, however, the average value is 7641 for imprints producing more than 100 editions (with a similar value for imprints publishing more than 500 editions). The higher concentration of format types for imprints implies that they tend to specialize in the format types they publish. Given this, it seems reasonable that they would also tend to specialize in the other physical qualities of their editions.

Electronic copy available at: https://ssrn.com/abstract=4547458

VI. Copyright and Quantity Sold

Although the main justification for copyright has been the expectation that it enhances the creation of new works, there have been some theoretic suggestions (with anecdotal evidence) that copyright also may induce post-creation investment in the title to keep the "franchise" going at an efficient level (Landes and Posner, 2003, Adilov and Waldman, 2013). The argument has been that the ownership provided by copyright allows a publisher of a title to eliminate the likely free riding on post-creation investment that would be expected for a title in the public domain (also inhibiting behavior that would decrease franchise value). This ability to fully harvest the returns from post-creation investment should increase sales, surplus and profitability for CR titles. Both Reimers and Heald, however, believe that they have tested this theory and found it wanting.[29]

Largely absent in these discussions, however, is the question of whether the fruits of greater post-creation investment would be found in the extensive margin of how many titles are in the market (as Heald presumed) or the intensive margin related to the sales levels of the titles already in the market. With current POD printing technology and the ability to say that editions are "in print" at almost zero cost, as seen in the previously noted entrance of tiny publishers with enormous listings of "phantom" editions, title "availability" as measured by BiP might not be very meaningful. It would seem better to measure availability by the actual exchange of editions in the market, as done in Section VII.

In contrast, post-creation investment might have its greatest impact on the intensive margin, the sales *quantity* of a title since most publishers do not just make copies of titles available and hope that some sell. Instead, they have sales agents and marketing departments to help get titles into bookstores, on to the better locations in bookstores, and into the consciousness of potential consumers. We turn now to the sales of titles.

A. *Measuring the Sales Difference*

To determine whether copyright affects the sales of titles we use BookScan's sales data for our vintage titles.[30] A comparison of CR to PD sales in 2004 is found in Table 3. It shows a remarkably large sales advantage for copyrighted titles relative to public domain titles. The third row of Table 3 indicates that the average copyrighted title sells almost 4 times as many copies as the average public domain title, and the median CR title sells more than 23 times as many units as the median PD title. These are exceptionally large differences.

[29] Heald believes that his finding that CR vintage titles are not more abundant than their PD counterparts is evidence against this post-creation investment theory, and Reimers believes that a coefficient value in her estimation argues against this hypothesis.

[30] Although Reimers had access to the BookScan data set, she only uses it to gauge the relative size of her "measured" sales of vintage titles on Amazon compared to sales of those titles in the overall market as measured by BookScan.

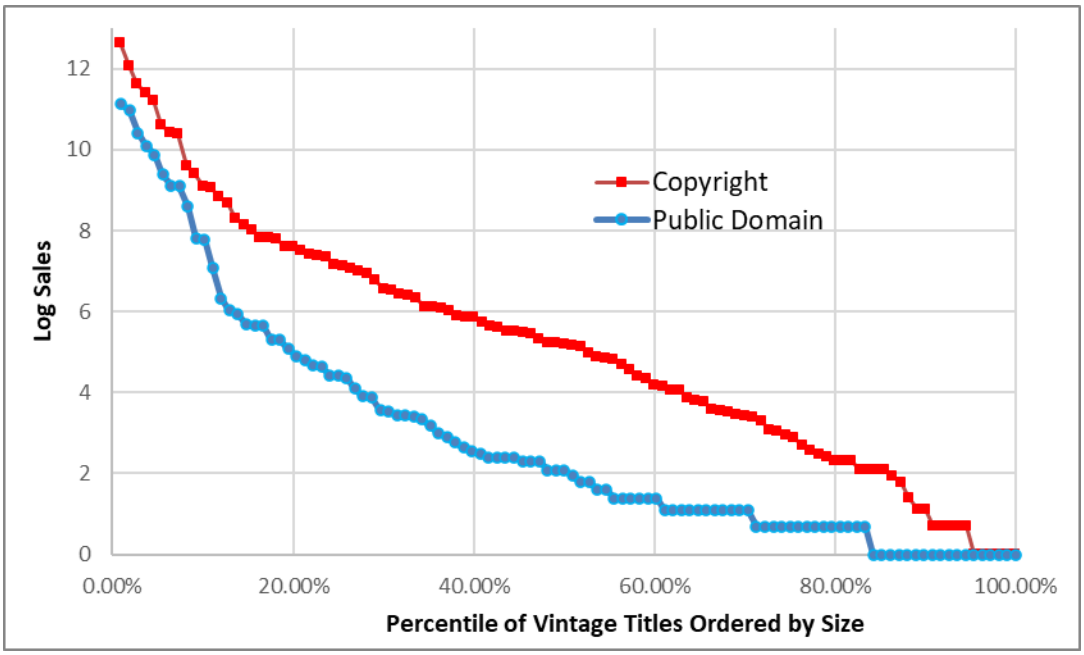
Electronic copy available at: https://ssrn.com/abstract=4547458

Table 3: 2004 Title Sales by Copyright Status

	obs	Mean sales	Median sales
Copyright	110	8,799	178
Public Domain	108	2,320	7.5
Ratio CR to PD		3.8	23.7

For anyone expecting copyright to embody textbook monopoly behavior of lower sales, higher prices, and the attendant deadweight losses, the numbers in Table 3 should be quite shocking. If these numbers accurately reflect the impact of copyright, something other than a copyright-induced exercise of monopoly power must be occurring in this market.

Figure 1: Title Unit Sales by Sample Percentile, from Largest to Smallest, 2004



One question left open by the numbers in Table 3 is whether the larger average sales for CR titles could be due to just one or two top selling titles, since sales of vintage best-sellers tend to be dominated by a relatively small number of top sellers, similar to the overall book market.[31] To determine how CR sales differ from PD sales throughout the samples, Figure 1 compares the (logged) unit sales for each PD and CR title with the same ranking percentile.[32] Because the number of PD and CR titles is almost the same (110 CR, 108 PD), the figure effectively compares titles one by one. The figure makes clear the rather remarkable result that each and every ranked

[31] The top 5% of vintage titles generate 84% of unit sales whereas for the full sample of 2004 titles the top 5% generate 86% of sales.

[32] Thus, if there were 100 PD titles and 50 CR titles, the 20th percentile would compare the unit sales of the 20th PD title and the 10th CR title.

Electronic copy available at: https://ssrn.com/abstract=4547458

CR title outsells its PD counterpart throughout the distributions, until sales converge on 1 unit [ln(1)=0]. Thus, all CR vintage titles sell better than their PD vintage title counterparts. In fact, the greatest CR sales advantage (in logged terms) is in the range of the 40th to 60th percentile, where CR titles tend to outsell PD titles by about 20:1. That is why the difference in median values is so large in Table 3.

We know that the distribution of sales for both PD and CR titles is not normal because they are highly skewed, have a very large range, and are greatly impacted by a small number of outlying values. Nevertheless, because each sample has slightly above 100 observations, parametric estimation should be viable, but because of the nonnormality we will also provide nonparametric testing of the sales differences between the samples. Using logged sales, a t-test comparing means for CR and PD titles has a value of 5.66. Similarly, the nonparametric Mann-Whitney U test, which is based on ranks, provides a significant z value of 5.63.

The unlogged raw data provide less statistical significance for CR titles' greater sales, with the t-test value being a statistically marginal 1.79, although the lower precision is not surprising when there are only a handful of titles that dominate sales, as seen in Figure 3 below. But this analysis is for 2004 alone. In Appendix 1, I reproduce this analysis for the years 2008, 2012, and 2016 and find results very similar those in Figure 1.[33] When the data for these four years are pooled, the t-test for differences in unlogged sales increases to 3.3, for logged sales it increases to 7.6, and the Mann-Whitney test provides a z-score of 7.8. With this more complete data the difference in sales, whether logged or not, is significant beyond the 1% level.

B. *Examining Potential Threats to Identification*

Before we can accept these sales differentials as evidence that copyright increases sales, however, we need to address the two potential identification problems mentioned earlier: competition from freely available digital copies of PD titles[34] and the greater recency of CR titles.

Although copyright precludes competition from nonprofit producers of free digital downloads, such competition is legal for PD titles. Free digital downloads would be expected to reduce sales of PD physical copies to some extent, but whether the impact is large or small depends on how willing consumers are to replace a purchased physical copy of a title with a free digital version.

In 2004, when our data begin, we would not expect free digital copies to replace the sales of physical books in any meaningful way because the reading experience from digital books was very inferior to reading hardcopy books. Any free digital titles in 2004 were going to be read

[33] For each of these years every CR title sells the same or more than the similarly ranked PD title, as in Figure 1.

[34] Reimers assumed that competition from free digital books fully explained the lower PD sales relative to CR sales. She states on page 263 "editions of protected [copyrighted) titles are sold more often on Amazon than editions of public domain titles, likely due to zero-price competition from Project Gutenberg and Google Books."

Electronic copy available at: https://ssrn.com/abstract=4547458

mainly on CRT desktop computer screens[35] (unless they were printed out, which would lessen or eliminate any cost advantage relative to physical books). Further, the Google book project was not announced until October 2004 so any free copies in 2004 were likely to come from Project Gutenberg, and its free digital files were created by volunteers who were likely to have less expertise to properly transcribe the text than traditional for-profit book publishers. The free digital download reading experience in 2004, therefore, would seem to be vastly inferior to using a physical book in terms of portability, eye strain, and text accuracy (typos and so forth). Thus, the large CR/PD sales differences in 2004 seem unlikely to be mainly due to free digital downloads.

But the changes in the market for digital books during the period of our data, 2004-2016, provides an ideal testbed to gauge the extent to which free digital PD editions might reduce sales of physical editions of PD titles relative to physical CR titles. During the 2004-2016 period, the digital reading experience vastly improved as new devices, such as the Kindle eReader and smartphones were launched (both in 2007). This much improved digital reading experience would be expected to increase the unit share of commercial digital titles (eBooks), as the first row of Table 4 shows to be the case, going from zero to the mid to upper teens.[36] The market share of eBooks can be considered a proxy for the ease and familiarity users are likely to have with downloading digital books. Because digital eBooks became better substitutes with physical books over this period, the competing-with-free hypothesis implies that the ratio of CR/PD sales should increase during this period.

Table 4: Change in Ratio of CR to PD Physical Title Sales Over Time

	2004	2008	2012	2016
eBook Share of Trade Units	0%	<1%	19.6%	15.0%
CR/PD Average sales	3.79	3.11	2.68	3.16
CR/PD Median Sales	23.73	7.37	6.76	5.05

Contrary to the predictions of the competing-with-free hypothesis, however, the second and third row of Table 4, which displays the ratio of CR to PD sales at four-year intervals starting in 2004, show that the CR/PD sales ratio (whether median or average) does not increase after the growth in familiarity and utility from digital books after 2008. The ratios of CR to PD sales in 2012 and 2016, when digital sales were large, are slightly lower than the values in 2004 and 2008 when

[35] Desktop computers, whose share fell over time, made up nearly three quarters of the flow of new personal computers in 2004 and thus had an even greater share of the stock of personal computers in use according to https://www.zdnet.com/article/73-3-of-pcs-sold-in-2004-were-desktops-24-laptops/. Cathode ray tube monitors outsold LCD screens until 2003 so that the majority of the stock of screens in 2004 would have been CRT screens.

[36] These values come from AAP StatShot.

Electronic copy available at: https://ssrn.com/abstract=4547458

consumers had almost no experience with eBooks. The large (and admittedly unexplained) decline in median values after 2004 is also contrary to this hypothesis.[37]

This failure to find support for the free digital book substitution thesis does not mean that there might not be a small effect. But it appears that we can rule out this "competing with free" thesis as a major explanation of the sales advantage of CR relative to PD titles. I conclude that competition from free digital downloads is not an important cause of the large CR/PD sales difference.

Why might free digital copies not greatly reduce sales of PD physical editions? Although only suggestions, here are some possibilities. First, perhaps most readers of the old-fashioned classic works in our sample tend to prefer old fashioned physical books as opposed to digital copies. Second, people frequenting sites for free digital titles might be switching from libraries as their next best alternative, instead of being former purchasers of physical copies. Third, the free editions may not have made their availability clear to prospective readers by, for example, not paying for (advertising) placements in Google searches.

We next turn to the suggestion that newer vintage titles will sell better than older vintage titles merely because they are newer (even though the newest vintage titles would have been 54 years old in 2004), causing CR titles to sell more than PD titles because CR titles are newer. While it is clearly true that the recency of publication matters a great deal in the sales of the title near the date of publication, as the sales of current bestsellers usually peak within a year or two of publication, it is less clear that this relationship would hold after a handful of decades have gone by.

To avoid conflating the impact of copyright from the impact of recency, we must examine the relationship between recency and sales separately for CR and PD titles. We begin by examining coefficients from the regression of sales on publication year, found in Table 5, for CR and PD titles together and separately for each of 2004, 2008, 2012, and 2016. The first pair of columns pool CR and PD titles together to reveal a significant positive relationship between logged sales and title publication year in all four 21st century years. The results from using unlogged values are also positive although not statistically significant (but approaching borderline significance in 2004). But it is unclear whether this relationship due to copyright or recency.

[37] In Figure 1 the distance between the two curves at the 50% level is large in 2004. In later years (as shown in Appendix 1) the sales of PD titles in that range of the distribution grow closer to the CR titles, but since the top 10% of titles are responsible for more than 90% of sales, the lack of change in that portion of the two distributions keeps the average values close to their original levels. Note that although the distance between median values gets smaller over time, they merely reach the same relative sales levels as the average sales. Importantly, the free digital substitute hypothesis implies that the PD sales would *fall* relative to the CR sales, not rise as they actually do.

Electronic copy available at: https://ssrn.com/abstract=4547458

Table 5: Coefficients of Sales on Publication Year for Vintage Titles

	Full Sample		PD (1895-1922)		CR (1923-1950)	
logged	Year Coef	t-stat	Year Coef	t-stat	Year Coef	t-stat
2004	0.074	5.84	0.02	0.65	0.087	2.42
2008	0.046	3.23	0.016	0.48	0.004	0.10
2012	0.047	3.00	0.013	0.33	0.020	0.54
2016	0.03	1.88	-0.037	0.94	0.034	0.95
unlogged						
2004	149.9	1.65	-98.2	0.97	-86.3	0.28
2008	143.3	1.51	-106.6	0.93	134.1	0.62
2012	87.8	1.17	-106.8	1.14	100.4	0.79
2016	118.4	1.20	-139.6	1.63	108.3	0.71

The next two pairs of columns show results for public domain and copyrighted works separately. There is no evidence that recency of publication is positively related to sales for those titles published during the PD period. The sign of the relationship is more likely to be negative than positive, and the coefficients are not significant in any year, whether we measure sales logged or not. We conclude that the sales of titles published in the 1895-1922 peroid are not affected by recency.

Nevertheless, the nature of a relationship between recency and sales is slightly less clearcut for titles published in the 1923-1950 period, as found in the last pair of columns. Although there is a significant positive relationship between sales and recency of publication for logged sales in 2004, the other three years of data do not show a relationship remotely close to statistical significance and using unlogged sales indicates a negative relationship in 2004. Further, in the three later years when there is no evidence of an impact of recency, there is still a large sales advantage for CR titles (as seen in Table 4). We conclude, therefore, that preponderance of evidence supports the view that copyright is responsible for larger sales separate from any impacts of recency.

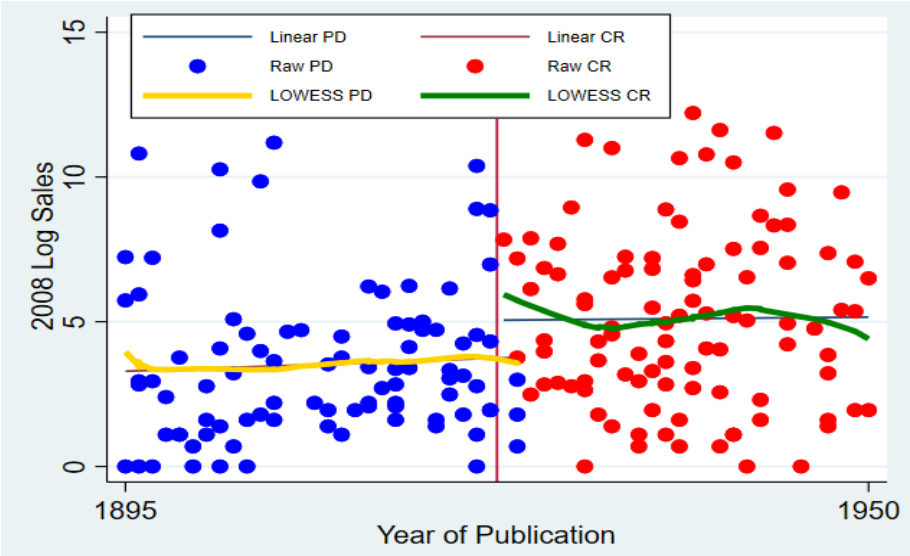
These conclusions can be put in visual form. Figure 2 shows logged sales and publication year for vintage titles in 2008, with the 1922/23 cutoff shown by the vertical line in the middle.[38] Figure 3 present the identical data but using unlogged numbers and a simple glance reveals the outsize influence that the few largest sellers are likely to have on industry measurements. On the left half of Figure 2 and 3 are the PD titles in blue, and on the right half are the CR titles in red. Both Figure 2 and 3 include LOWESS smoothed curves, green for CR titles and gold for PD titles, as well

[38] The three public domain points that are mixed in with points representing copyrighted works are titles published after 1922 that entered the public domain because they were not renewed 28 years after publication.

Electronic copy available at: https://ssrn.com/abstract=4547458

as linear regression lines.[39] I use 2008 data in Figures 2 and 3 because 2008 is somewhat more representative of typical results than are the 2004 data, as seen in Table 5.

Figure 2: Titles with Logged 2008 sales and Publication Year



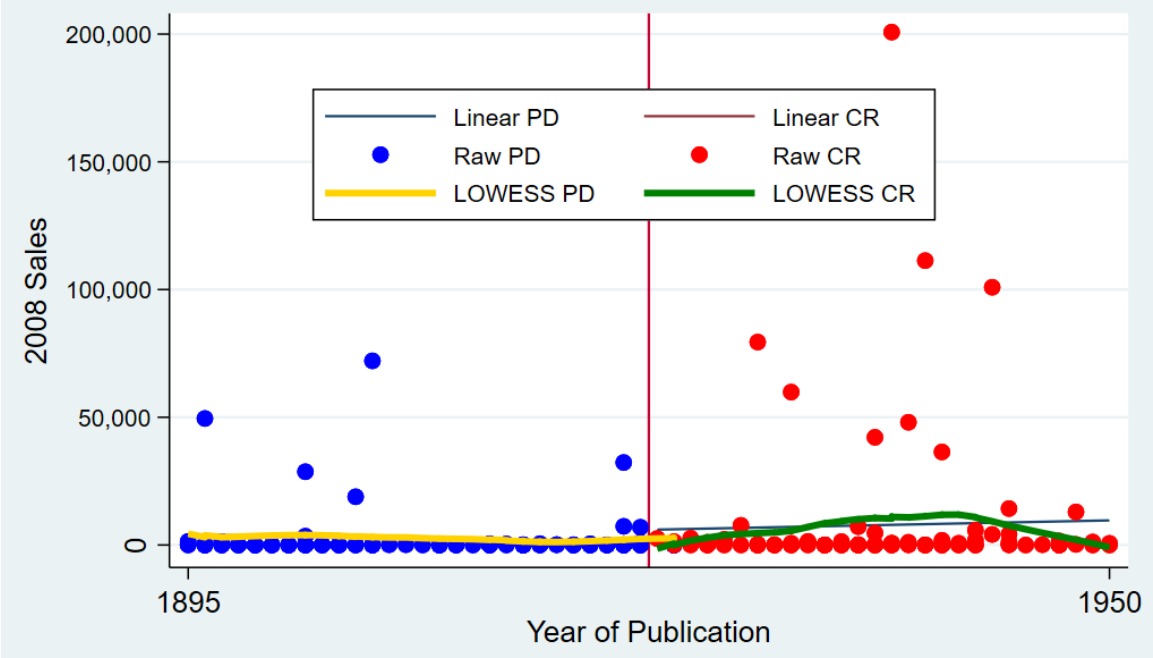
The PD LOWESS curve in Figure 2 rises slightly whereas in Figure 3 it falls slightly, consistent with Table 5. Neither curve shows much curvature. For CR titles with logged values the LOWESS curve falls, rises, and falls again in Figure 2 (a pattern replicated in the other years), whereas in Figure 3 it rises and then falls, finishing about where it started. The decline in the last decade is particularly at variance with the recency hypothesis. Equivalent figures for other years are found in Appendix 2.

Another approach to determine whether the greater recency of copyrighted titles leads to higher sales of copyrighted titles is to weaken the possible impact of recency on sales. This can be done by shortening the interval between the oldest and newest vintage titles which weakens any potential impact of recency since the most recent titles will be not much newer than the oldest titles.

[39] The bandwidth parameter used in the LOWESS smoothing is 0.8. I also tried smaller bandwidths (down to .1) to make the curve more localized, but it did not change the basic relationship of the two lines.

Electronic copy available at: https://ssrn.com/abstract=4547458

Figure 3: Title Sales by Publication Year



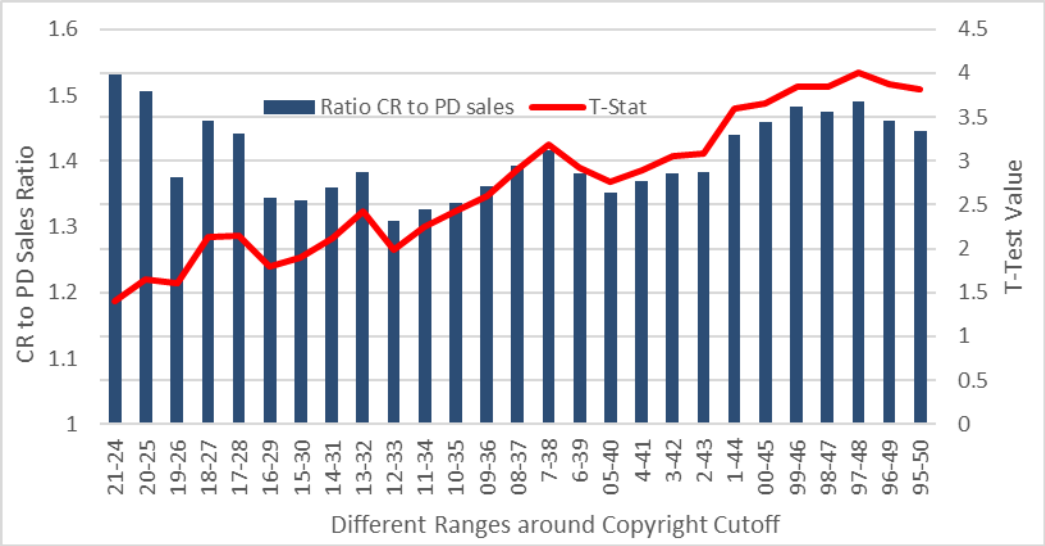
For the full sample, the oldest titles are 56 years older than the newest titles. But if we reduce the sample by some amount, half (three quarters) for example, removing an equal number of the oldest and youngest titles when reducing the sample, the range will drop from 56 to 28 (14) years, reducing the ability of recency to affect results. A secondary advantage of this approach is that it focuses on titles that are close in age and therefore less likely to be differentially affected by unobserved long run forces other than recency. Unfortunately, doing so also entails reducing the size of the sample.

Figure 4 illustrates some of these results for logged values in the year 2008. It shows the CR/PD sales ratio for different intervals around the cutoff, represented by the dark bars. Since the bars are always greater than 1, CR sales are greater than PD sales for every range around the copyright cutoff. The continuous red line represents the (robust) t-statistics whose values (and grid lines) are on the right-hand axis. The sales advantage for CR titles is statistically significant for all ranges with more than 9 years on each side of the cutoff, but it is also significant (at the 5% level) when there are only 5 or 6 years on each side of the cutoff and is always at least of borderline significance when there are more than 2 years on each side of the cutoff. The sales advantage of CR titles using logged values (which tend to range only between 1.2 and 1.5) may appear to be small due to using logged values, but the differences are much larger if we exponentiate the logged values.[40]

[40] If we exponentiate the numerator and denominator of the average logged values forming these ratios, the resulting ratios range from a low of about 2 to a high of about 8.

Electronic copy available at: https://ssrn.com/abstract=4547458

Figure 4: 2008 Comparison of CR and PD Log Sales Using Various Ranges around Cutoff



Appendix 3 provides similar results for the years 2004, 2012 and 2016, plus results for each year using unlogged data. For the most part the already noted pattern of CR sales being considerably larger than PD sales is generally robust to decreases in the range around the cutoff. This is always true for logged sales, and mostly true for unlogged sales.[41] It is also generally the case that the differences between CR and PD logged sales are statistically significant over most of the possible yearly ranges around the copyright cutoff, even in those cases where the number of observations is quite small. As was the case for the full sample, the sales difference using unlogged data generally does not (but occasionally does) reach normal levels of statistical significance. The results are strongest in the early years of our data (e.g., 2004) and weaker in the later years (e.g., 2016), but the overall results are consistent with a view that reducing the effects of recency by reducing the number of years around the copyright cutoff provides results consistent with our other findings.

The conclusion of this section is that copyright increases the sales of titles. The two potential alternative explanations, competition with free digital downloads and the recency of copyrighted works, have been examined and found to be unable to account for the large advantage in sales for CR works. This finding of copyright causing greater sales is consistent with a strong post-creation investment effect. Admittedly, this is only indirect evidence for post creation investment because we are not measuring post creation investment directly. Nevertheless, it is a positive sales effect that is brought about by some aspect of copyright, and it is not clear that it matters what aspect of copyright brings about the additional sales, although the post-creation investment hypothesis must be a leading candidate.

[41] For unlogged sales which are influenced by its outlier values, it is not until the first CR outlier occurs, in the 1916-1929 range, that CR sales become larger than PD sales. The CR sales advantage continues for all larger ranges. This pattern holds for each of the years.

Electronic copy available at: https://ssrn.com/abstract=4547458

VII. The economic importance of top vintage titles

It is useful, when considering publishers' use of resources in post-creation investment, to understand that although these vintage titles do not have yearly sales that are at the top of the current yearly sales distributions, these titles can nevertheless be of considerable financial consequence to publishers because of the durability of their sales. Some of these very old titles continue to sell in large enough quantities year after year so that in a timeframe measured in decades, or even a century, they would often surpass, sometimes by a wide margin, many current best-sellers' lifetime sales (except those destined to become longtime classics themselves). These vintage titles are like the apocryphal tortoise that can outdistance the hare, so to speak.

Over the 13-year period of our sales data, the top 100 bestsellers had 13-year sales that ran from about 1.2 million units to about 6.4 million (a single outlier sold 9.2 million), with the 20 top titles each selling more than 3 million units and the top 50 each selling more than 1.9 million. In that same period, the top selling vintage title sold about 2.5 million titles, and the top 5 vintage titles averaged 1.2 million units, with a median value of 804 thousand.[42]

Back-of-the-envelope extrapolations of these results to a longer timeframe, say 52 years, would lead to sales for our vintage titles that are four times as high as found in our 13-year period, ignoring population growth and other changes. Extrapolating values in this manner makes it clear that these top selling vintage titles are very valuable properties indeed. The top vintage title, with expected sales of about 10 million units sold in 52 years, would surpass any bestseller in our 13-year period, and the other leading vintage titles would be among the very top current sellers. For longer time periods, these leading vintage sellers' sales would be so large that they clearly would justify considerable post-creation investments.[43]

Even the vintage titles that are somewhat below these top sellers have the financial benefit of a very large number of years with sales, improving their lifetime sales well above what a typical title with similar yearly sales might achieve. Therefore, it is not just the top vintage sellers that have sufficient market potential to warrant significant post-creation investments, but also more moderate sellers that might not otherwise be thought to justify such investment given the yearly sales of these titles.

VIII. Factors that affect the Cost and Price of Books

One practically universal expectation is that copyright will affect the price of books. To explore that expectation, we need to examine factors that affect the production cost of books. The most

[42] Some additional factoids are that the top 2004-16 selling title from 1895-1950 (not among our vintage bestsellers) sold 3.9 million copies and the top old title (1960) was number 2 in 2004-16 overall sales, selling 6.4 million copies.

[43] Even though these classics might sell the most total copies, we would need to know production costs and present values of these numbers to determine how these classic titles compare to traditional bestsellers in terms of profitability.

Electronic copy available at: https://ssrn.com/abstract=4547458

obvious factor is the type of physical format. Another factor that will obviously influence the cost of production is the number of pages.

Illustrations in books are another additional expense, both in terms of paying the illustrator and for the additional printing costs, particularly if the illustrations are in color. The relative share of pages taken up by illustrations would be an ideal measure, but we are limited to a dummy variable indicating whether or not the book has illustrations.[44]

Other factors, such as paper size and weight, ink quality, dust jackets, cover art, and so forth also affect the cost of book production, but data at this level of detail is in general not available, although it seems likely that imprints, and to a lesser extent publishers, will tend specialize somewhat in the general quality books they produce (see the discussion around footnote 28).

An additional possible factor that might influence the price of an edition is the age of the edition, measured as years since publication of that edition (not the publication of the title). In 2004, half the fiction editions sold were more than 3 years old, a quarter of editions were more than 7 years old, and one tenth of the editions were more than 12 years old. Because there was modest inflation over the decade prior to 2004, the later the year of an edition's publication, the higher the nominal price would be expected to be, ceteris paribus.

The final factor examined is economies of scale. Given the nature of printing, economies are to be expected. The reasons are, first, that when the scale of printing increases sufficiently, publishers are likely to switch from the relatively expensive POD to lower cost offset (lithography) printing, and second, larger print runs decrease the average cost per unit when using offset printing. Appendix 4 examines evidence of economies of scale in book production. Although the evidence is consistent with economies of scale for hardcover and trade paper editions, it is inconsistent for MM paperbacks for reasons not yet understood.

We also examine how editions from major publishers differ from editions printed by minor publishers. Although the evidence of economies of scale is somewhat mixed, we proxy for larger print runs in the regressions by whether an edition is published by a major publisher. Editions sold by major publishers greatly outsell editions from minor publishers in the industry as a whole, with the average and median yearly sales of the former (2,974; 133) being much larger than the sales of the latter (235; 7). This difference is even greater for our sample of vintage titles, as seen in Appendix 4.

Table 6 shows how these cost factors are related to the price of editions, first for the entire market and then separately for each type of format. Clerides (2002), using data from a leading academic publisher, concludes that book prices were affected by cost shifters, not demand factors. If this is generally the case, regressing the log of price of an edition on the four cost factors (number of pages, edition year, major publisher dummy, and illustration dummy) and the format dummies,

[44] The information on the inclusion of illustrations came from the WorldCat database.

Electronic copy available at: https://ssrn.com/abstract=4547458

should provide realistic estimates. We also remove collectibles (editions priced over $65) and control for imprint fixed effects. Virtually all of these editions are copyrighted.

Table 6: Edition Characteristics influencing log price, 2004

	all	No Imprint FE	hard	paper	MM
Pages (x100)	0.063	0.065	0.040	0.069	0.095
	(66.23)	(72.05)	(29.97)	(50.09)	(53.32)
Edition year (x10)	0.0756	0.0715	0.0531	0.0939	0.0648
	(31.22)	(33.98)	(14.84)	(27.84)	(17.74)
Majorpub	-0.0379	-0.04478	-0.06077	-0.03381	-0.02226
	(-5.15)	(-20.77)	(-18.61)	(-10.65)	(-4.05)
Illustration	0.0310	0.02206	0.05859	-0.00006	0.05571
	(8.45)	(6.81)	(8.7)	(-0.02)	(5.85)
format fixed effects	x	x			
remove collectables	x	x	x	x	x
Imprint fixed effect	x				
N	107,263	107,960	25,649	54,104	28,207
adj. R-sq	0.85	0.72	0.08	0.10	0.19
Dep Var= Log Price; robust t-stat in parenthesis; titles with Price>65 removed.					

As seen in the first column, all variables have the expected sign and achieve statistical significance. The second column is the same except for the fact that it excludes the imprint fixed effects. The results are similar with or without fixed effects. The next three columns show the results for each format type separately, with very similar results. The only surprising result is that illustrations do not appear to raise the price of trade paper editions.

IX. The Impact of Copyright on Price

If we wish to know the impact of copyright on price for our sample of vintage titles, we need to focus on the price controlled by the copyright owner. The copyright owner or their assignee (usually the publisher) determines the wholesale price charged to a retailer and also the suggested retail price, so those would be the prices to examine if they were available.[45] Because

[45] The list prices do not change over the life of the barcode representing that ISBN, although a price can change for the ISBN if the barcode is changed, although that seems to happen rarely if at all. See the discussion of barcodes at Bowker's FAQs https://www.myidentifiers.com/faq/barcodes. I looked for any price changes over the 13 years of my data for approximately a million editions and discovered only 241.

Electronic copy available at: https://ssrn.com/abstract=4547458

we do not have access to the wholesale price charged to a retailer, we will use the retail list price.[46]

As a practical matter, retailers can charge whatever price they wish although many retail bookstores typically charge the list price that is found on the book cover, but with large discounts for current bestsellers being common. Author royalties are also normally based directly on the list price, although occasionally they are based on the wholesale price (net cash received). Most previous researchers have used list prices when comparing prices for copyrighted and public domain works, although Reimers (2018) uses Amazon retail prices.

Table 2 in Section IV provided some simple price comparisons. We saw that CR vintage editions had 16% ($20.63/$17.86) higher average prices than PD editions in 2004. We also saw that CR editions were more likely to be hardcover and to have more pages, implying that the higher copyright price might be due to these factors. On the other side of the ledger, CR editions were more likely to be published by major publishers, and be older (note: not the title, but a particular edition), each of which tends to lower price.

Table 7 examines the impact of copyright on price, controlling for factors found to alter price in the market, but for the sample of vintage titles. The first column represents the average impact of copyright, with copyrighted editions selling at a premium in the vicinity of 20% relative to editions in the public domain. This premium is what other researchers have attempted to measure. All the control variables have the expected sign, and they are statistically significant. For simplicity, we exclude illustrations because its coefficient is always economically and statistically insignificant.

Table 7: 2004 Regression Results Explaining Price of Vintage Editions

Log Price	Full	top quartile	2nd quartile	3rd quartile	Bottom quartile
Copyright	0.202	0.508	0.169	0.185	0.086
	(5.49)	(8.72)	(2.52)	(2.17)	(0.90)
Pages (x100)	0.0846	0.073	0.075	0.082	0.0810
	(6.59)	(3.73)	(2.90)	(3.07)	(3.00)
Majorpub	-0.192	0.060	0.038	-0.066	0.015
	(-4.55)	(0.83)	(0.50)	(-0.43)	(2.19)
Edition Year	0.0743	-0.002	-0.004	0.015	-0.1110
	(3.84)	(-0.85)	(-1.21)	4.23	-0.85
Observations	533	136	136	129	132
Adj R sq	0.53	0.68	0.54	0.53	0.36

Log price is the dependent variable, t statistics in parentheses, robust SE, P<65, Sales>0, format fixed effects.

[46] Typical trade discounts (off list price) are currently about 55% https://www.ingramspark.com/blog/why-should-i-discount-my-book.

Electronic copy available at: https://ssrn.com/abstract=4547458

The copyright premium found in the first column, for the entire sample, appears to be somewhat lower than the prior estimates discussed in the literature review. But the methodologies used in Heald or Li, MacGarvie, and Moser are so different that it is not clear we are measuring the same thing. Reimers' estimate (a coefficient of 0.24 leading to a 27% premium) appears to be only slightly higher than that found here. In fact, however, as seen in Appendix 5, when used books (which make up 22% off her sample of editions) are excluded, as they should be,[47] and the number of pages is included as a control, the estimated copyright coefficient drops to a statistically insignificant 0.07. Therefore, the estimate in the first column is estimate is actually considerably higher than what comes from her data.[48]

It is somewhat difficult to put this ~20% figure in perspective. It would be helpful to know what the typical royalty rates were during the era of our vintage titles publications. Historical American author royalties in the mid-19th century were generally in the vicinity of 10% but reached as high as 50% for famous authors such as James Fenimore Cooper or Washington Irving (Liebowitz, 2016a). According to industry insiders, current royalties to authors top out at 15% for hardcover sales above 10,000 units, 10% for paperbacks (over 150,000 units), and 25% of net receipts for eBooks but could be almost twice those levels if publishers were not giving such large advances.[49] Because our sample consists of bestsellers mainly by well-known authors, we would expect their royalties to be at the highest end of the distribution. The average copyright premiums implied by our results, therefore, seem to be well within the general vicinity of likely royalty rates.

But it is not clear that we should be focused on the *average* copyright price differential. The next four columns of Table 7 show the same specifications for subsamples of the data based on unit sales of editions. What these subsamples tell us is that the impact of copyright on price is considerably larger for top selling editions than it is for poorer or average selling editions.[50] Further, the editions in the top quartile are responsible for 98.9% of the full sample's sales. That would mean that consumers pay an average premium of about 50%, not 20%, for the copyrighted editions that they purchase. This 50% value seems very likely to be higher than the royalty rates being paid by publishers, although it is not clear whether they would be higher than the royalty plus the cost of providing post-creation investment.

---

[47] Used books do not belong in the sample because the first sale doctrine (Mortimer, 2007) prevents copyright owners from generating any revenues from or having any control over the used book market.

[48] Her failure to include number of pages is surprising because in her footnote 45 she states: "the estimated marginal costs are strongly positively correlated with the number of pages in the edition, conditional on the edition's format." Her estimated copyright coefficient drops to 2.5% if number of pages alone is added to her specification.

[49] See a post from an industry insider https://www.thepassivevoice.com/are-royalties-fair-a-publisher-weighs-in/ who reports the standard royalty rates but also claims "royalties could be roughly 70% higher in a world without advances."

[50] These results also tell us that the overall impact of the major publisher dummy must have something to do with the composition of titles produced by major publishers since its relationship to price is very different in each of the sales quartiles compared to the pooled sample.

Electronic copy available at: https://ssrn.com/abstract=4547458

It is natural to suspect that publishers of the most successful copyrighted vintage titles are taking advantage of their position as sole sellers of popular copyrighted titles to charge these higher prices. But if that were all that was going on, then CR titles should not outsell successful PD titles the way that they do, particularly at the high sales portions of the sales distribution. The higher price for popular CR titles, ceteris paribus, would be expected to lower the sales of top selling CR titles relative to top selling PD titles. Thus, an explanation more consistent with the facts that we have found would be that some or all of the revenues from the higher prices for leading CR titles, in addition to paying royalties, would seem likely to go to producer activities that help sell these better selling editions.

X. Copyright and "Availability" of Vintage Titles

One of the questions of interest in this literature is whether a copyright regime increases or decreases the availability of already published titles. Heald concludes that physical editions of PD titles, in more recent years, are just as if not more likely to be made available in the market (as evidenced by BiP) than are copyrighted works. He believes this runs counter to the expectation that competition could reduce the market viability of some titles which might only survive under monopoly pricing. Looking at editions, not titles, Reimers and Heald both conclude that copyright reduces the number of editions available per title.

A question they did not answer is whether the greater availability indicated by BiP data translated into greater consumption variety. Heald understands the problem, when he says: "Arguably, comparative sales data would be a better measure of availability than in-print status; however, historical sales data are generally not publicly available" [p. 1038]. Since we have such sales data, we can measure consumption availability.

Table 8: 2004 Availability of Vintage Bestselling Titles

	Total (1895-1950	Copyrighted (1923-50)	PD (1895-1922	t-test CR = PD
1. Yearly Avg # New Vintage Bestsellers	8.80	8.63	8.96	
2. % Vintage Titles "in-print" (BiP)	65.0%	54.5%	75.6%	5.12
3. % Vintage Titles with Sales	44.3%	47.5%	41.1%	1.37

Table 8 provides the numbers for this latter type of "availability" in 2004. To begin, we need to determine the number of new bestsellers that were published each year during the 1895-1950 period. As already explained, my data consist of the top ten bestsellers for each year between 1895 and 1950 so one might think this would lead to 10 being the answer. But because a bestselling title in one year is often also a bestselling title in a later year, the actual number of first-time vintage bestselling titles in a year is generally less than 10, and this number is shown in row 1 of Table 8.[51] The number of possible first-time bestsellers is somewhat higher for PD titles,

[51] In two years (1898 and 1922) there was a tie for the tenth-place bestseller, so that 11 "top ten" titles were listed.

Electronic copy available at: https://ssrn.com/abstract=4547458

indicating that titles were slightly more likely to appear multiple times in the bestseller lists prior to 1923 than after, although this difference is obviously unrelated to current copyright status.

Using the hardcopy 2004 BiP, Row 2 shows that 65% of the potential titles in our sample were listed as in-print by BiP in 2004. PD titles were considerably more likely to be in-print than CR titles, with in-print values of 75.6% and 54.5% respectively, a statistically significant difference. This apparent finding that copyrighted vintage titles are less likely to be in-print than PD vintage titles is largely consistent with what Heald found.[52] The implication would seem to be that copyright decreases the availability of titles. This has been taken by Heald to support the view that the removal of copyright does not lead to the underexploitation of already created works.

The BookScan data set allows us to go directly to the share of vintage titles that are sold each year, as opposed to merely being in-print. There are two sources of differences between BiP and BookScan data. First, there are some BookScan listed editions that are either not in BiP or were not found in BiP.[53] Second, and probably more important, many BiP listed editions were found to not have any sales and thus do not appear in BookScan. The third row of Table 8 compares the shares of potential vintage titles that are sold in 2004, by copyright status. Contrary to the seemingly greater availability of PD titles in row 2, CR titles are somewhat *more* likely to be sold than are PD titles, although the latter difference is not statistically significant. This result is consistent with the claim that copyright's prevention of free riding would enhance the number of titles "available" to consumers although the result is too weak to verify that claim.[54] Of note to researchers in the field, the "availability" of PD titles relative to CR titles appears quite different depending on whether we use sales or in-print status to measure it.

In addition to comparing average availability, we can check the trends of availability to make sure that the difference in means is not due to some underlying time trend. Figure 5 shows, for each publication year, what share of the yearly vintage best-sellers are sold in 2004, with the 1923 copyright cutoff shown as the vertical line. The data points indicate no apparent visual trend for PD or CR titles (a regression indicates .09 extra titles per PD decade and -.04 fewer titles per CR decade, although neither is statistically significant). We conclude from this that the values in Table 8 do not mask important trends. There is also no evidence of a recency effect with respect to the availability of titles as measured by positive sales, a result consistent with our prior observation of the lack of recency affecting sales.

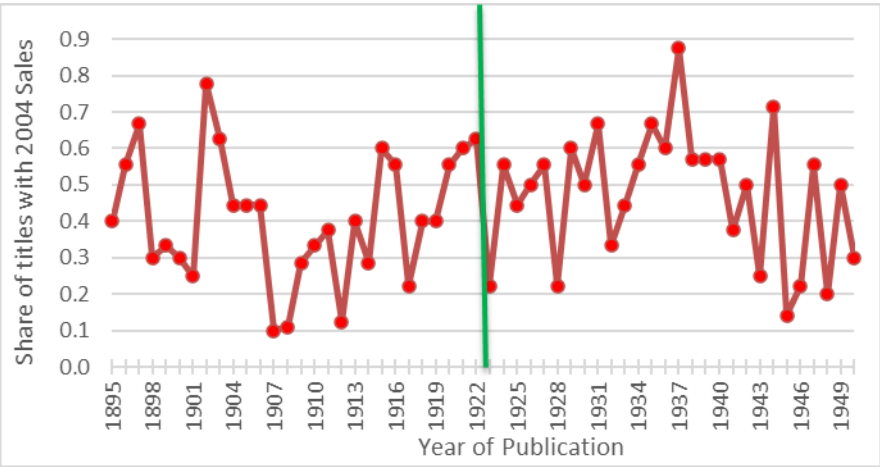
[52] Heald found that copyrighted and PD vintage works were about equally likely to be available in in the 1980s and 1990s editions of BiP (see his Figure 1). But by 2004, PD works were considerably more likely to be found in BiP (his Figure 1, whose values seemed to be about 55% and 86% for copyrighted and PD works respectively).

[53] BookScan has editions of titles, sometimes the highest selling editions, that Reimers did not find in the electronic version of BiP, as discussed in Appendix 5.

[54] None of this is to say that PD titles are not more likely to be available than CR titles *in some other form*, particularly since there are organizations such as Project Gutenberg or Google which make PD titles available as free digital downloads.

Electronic copy available at: https://ssrn.com/abstract=4547458

Figure 5: 2004 Share of Vintage Titles with Sales by Publication Year



Finally, both Reimers and Heald examine how copyright affects the availability of *editions per title*, although additional variants of a title would seem to provide far less novelty and value than additional titles, making the importance of more or fewer editions per title of questionable interest. Nevertheless, both researchers find that there are more editions for PD titles than for CR titles. Reimers finds that there are 43.7 editions for each PD title and only 7.4 editions for each CR title, an exceptionally large difference.[55] Heald limits editions to 1 per title for each publisher and finds a smaller advantage of 5.2 editions per PD title versus 3.2 editions for CR titles.

Table 9: Editions per Title by Copyright Status

	total	copyrighted	PD	T-stat
Editions per listed BiP title	2.2	2.0	2.4	1.63
Editions per title with sales	2.4	2.0	2.9	2.53

Our results, found in Table 9, have the same sign as found by Reimers and Heald. Although I find more editions of PD titles than CR titles, the difference is only 21% (2.4/2.0) compared to the 590% for Reimers or the 62% for Heald. When I limit the comparison to editions with sales, the difference between PD and CR increases to a somewhat larger 49% (2.9/2.0). Of course, it should not be surprising that PD titles have more editions sold when they are freely available for any publisher to sell.

XI. Implications for Welfare

How do these results fit into the conventional thinking about the welfare effects of copyright? In the most traditional view, copyright was supposed to reduce surplus from the consumption of copyrighted works, with the reduced consumer surplus being "balanced" by the increase in welfare from copyright's inducement to produce new creative works. A modification of this

[55] Found in the third row of her Table 1.

Electronic copy available at: https://ssrn.com/abstract=4547458

model suggests that copyright might benefit society after a work is created by promoting post-creation investment in a title, for which we have found indirect support in the higher sales of copyrighted works. Another modification can be made if one assumes that book production is a natural monopoly and that allowing competition in the production of multiple editions of a title may therefore be inefficient. But the main criticism of copyright has always been that it reduces sales.

The large positive impact of copyright on sales raises the possibility that the impact of copyright on the consumption of already created works might be positive. If this were the case, however, then the traditional welfare analysis of copyright would be stood on its head. The incentive/access relationship would no longer be a tradeoff since copyright would be socially beneficial on both sides of the 'balance,' and that would mean that copyright would clearly be beneficial to society (ignoring its effect on follow-on works).

But that is not all. The criticisms of retroactive copyright extensions (Akerlof et al., 2003) would be incorrect. Similarly, the supposed benefit to society from allowing already created works to fall into the public domain, which has been taken for granted by so many, would also be incorrect.

A welfare analysis of these issues was the main focus of Reimers' 2019 study. Reimers uses a discreet choice model, following Berry and Waldfogel (1999), to estimate consumer and producer surplus in a world with and without copyright. She concludes that copyright reduces welfare if its impact on the creation of new works is ignored. Her analysis relies on a self-created measure of sales based on changes in Amazon sales rankings, but this sales variable turns out to be inadequate for its task as discussed in Appendix 5.

I offer no judgment of her welfare methodology except to note that it is highly regarded within much of the profession. Instead, I treat her method as a black box to see what its results would be if better (BookScan) data were used as the measure of sales. Making just one simple data adjustment (replacing Reimer's sales values with BookScan sales values for the same titles) turns out to reverse her conclusions, as demonstrated in Appendix 5, which also considers other possible issues with Reimers' analysis.

Table 10: Average Results per Title for Vintage Titles using Reimer's Model

		Consumer Surplus	Profit	Total Surplus
(1) Reimers Original Results	CR	$2,550	$5,460	$8,010
	PD	$12,532	$0	$12,532
(2) Using BookScan Sales Data	CR	$6,287	$22,377	$28,664
	PD	$18,461	$0	$18,461

The first row in Table 10 replicates Reimers' estimate of the average surplus among her vintage titles for 12 months from September 2011 to August 2012. Total surplus (with the assumption that profit is equivalent to producer's surplus) is found in the right-most column. The main result

Electronic copy available at: https://ssrn.com/abstract=4547458

of her analysis is that the average surplus for public domain titles is about 50% larger than the average for copyrighted titles, at approximately $12,500 per PD title relative to $8000 per CR title. These are the values leading Reimers to conclude that copyright is generally harmful to social welfare when its impact on the creation of new works is ignored.

Row 2 replaces the sales data used by Reimers (based on counting each improvement in a title's Amazon sales ranking as a sale) with 2012 data from BookScan.[56] As I explain in Appendix 5, her method of counting sales has some deficiencies, particularly when an edition sells dozens of copies per day or more. This is a particular problem for a handful of top selling vintage titles which, according to BookScan, sell up to 650% more copies than her measurement of market sales for these titles.[57] Replacing her measured sales with the BookScan sales numbers dramatically alters and reverses her results. Row 2 of Table 10 indicates that replacing her sales values with BookScan values causes CR titles to have a surplus approximately 50% larger than the surplus generated by PD titles. In other words, using more accurate sales data reverses her results, making them consistent with our finding that copyright enhances the sales (and presumably the surplus) of titles on the access side of the market.[58]

What we can definitively say from this analysis is that Reimers' methods, ignoring the incentive impact of copyright, cannot be used to claim that PD titles generate greater average surplus than do CR titles. If one has faith in her methodology, however, it appears that a stronger statement can be made, which is that the access portion of the copyright balance is enhanced by copyright, contrary to her published results and to textbook expectations that copyright induces monopoly output restrictions. I am somewhat loath to accept this claim based on someone else's empirical modelling, but I am at least heartened by its general consistency with our earlier result that copyright seems likely to be socially beneficial independent of its impact on the incentive to create titles.

Conclusions

We have endeavored to use a natural experiment in copyright assignment to examine how copyright has altered the sales, pricing, and availability of vintage fiction titles. Our access to an

[56] Reimers estimates monthly sales data over 12 months and then multiplies each of her monthly values by 2.5 in an attempt to bring her yearly Amazon sales number into alignment with industry sales as measured by BookScan values. Although Reimers uses BookScan data to perform this extrapolation, she fails to make the correct adjustment and does not use the BookScan data to adjust her measured Amazon sales for individual titles. This is explained in more detail in Appendix 5.

[57] I double checked that the 2011 sales of the top titles were similar to the 2012 data to make sure that the slight mismatch in months covered in our data sets is not responsible for the results.

[58] I also examine Reimers' measurement of the value of free digital copies of PD titles, concluding that it is overstated and that correcting it can also reverse her conclusions, although CR surplus is only a little larger than PD surplus. But that analysis, found in Appendix 5, is not needed for my conclusion although it increases the advantage of CR titles beyond what is found in Table 10.

Electronic copy available at: https://ssrn.com/abstract=4547458

unusually rich data set has led to findings that call into question the standard understanding of the impacts of copyright.

We have discovered that copyrighted titles generally sell in much greater quantities than similarly situated public domain titles. The average difference in sales is almost four to one in favor of copyrighted titles, and the sales advantage holds throughout the full distribution of sales. We have tested alternative hypotheses that might have explained this result and found them lacking in support. The implication is that copyright does not cause a deadweight loss due to quantity restrictions, which is a remarkable result overturning most of what we think we know about copyright's welfare effects.

The most natural explanation for the greater sales of copyrighted works would seem to be post-creation investment by the publishers of titles, where publishers of copyrighted titles can make marketing investments with the knowledge that they will reap the full return on investment unlike the publishers of public domain titles who would share the returns with free riding competitors. This is the first evidence of which I am aware, other than case studies, supporting the post-creation investment hypothesis.

We also have three other findings. First, copyrighted titles have higher prices than public domain titles, although the average price differential is somewhat smaller than found in most other studies. Nevertheless, the price differential is considerably higher for the more successful titles comprising the lion's share of sales. These higher differentials are above most but not all royalty rates paid by publishers to leading authors. Thus, it is not clear whether the copyright premium is greater than the royalty costs plus post-creation investment costs, and that is a question for further research. Second, copyrighted vintage titles are slightly more likely to be sold in the market than are PD titles, contrary to claims based on "in-print" status which tend to find PD titles more available. One reason for this difference appears to be that new business models in publishing have led to serious differences between editions being listed as in-print and actually being sold in the market. Third, after inserting our superior sales data into a previously published study looking at the welfare implications of copyright we found that it reverses the conclusions of that study and matches our finding that copyright enhances welfare even if any incentive effects of copyright are ignored.

The most intriguing implications of these findings is the possibility that the access/incentive tradeoff may not be a tradeoff at all, at least not with respect to the works of fiction examined here. If copyright provides an incentive for authors to create new works, and if copyright increases the sales of already written works, then copyright would appear to be unambiguously beneficial. It would also mean that retroactive extensions of copyright, which have, in the main, been considered clearly harmful, might in fact also be beneficial. This result makes a strong case for indefinitely renewable copyright. It might also mean that the boundaries of fair use should be narrowed because the harms from fair use may be larger than previously thought if it interferes with post-creation investment.

Electronic copy available at: https://ssrn.com/abstract=4547458

Do these results hold for other types of books? Do these results hold for other types of copyrighted works or other types of intellectual property? Clearly, further work on these topics is called for if we are to understand how copyright affects consumption and production of intellectual creations.

Electronic copy available at: https://ssrn.com/abstract=4547458

AAP StatShot various years, Association of American Publishers, 455 Mass Ave, Suite 700, Washington DC 20001.

Adilov, Nodir, and Michael Waldman. 2013. "Optimal Copyright Length and Ex Post Investment: A Mickey Mouse Approach." *Economic Inquiry* 51 (2): 1101–22.

Berry, Steven T., and Joel Waldfogel. 1999. "Free Entry and Social Inefficiency in Radio Broadcasting." RAND Journal of Economics 30 (3): 397–420.

Brief of George A. Akerlof et al. as Amici Curiae in Support of Petitioners at 12, Eldred v. Ashcroft, 537 U.S. 186 (2003) (No. 01-618).

Boldrin, M and D. K. Levine *Against Intellectual Monopoly* Cambridge University Press, July 7, 2008.

Clerides, Sofronis K. "Book value: intertemporal pricing and quality discrimination in the US market for books" *International Journal of Industrial Organization* 20 (2002) 1385–1408.

Ellison, Glenn, and Sara Fisher Ellison. "Match quality, search, and the Internet market for used books." No. w24197. National Bureau of Economic Research, 2018.

Heald, Paul J., "Property Rights and the Efficient Exploitation of Copyrighted Works: An Empirical Analysis of Public Domain and Copyrighted Fiction Bestsellers," *Minnesota Law Review*, April 2008 92(4) 1031-1063.

Kitch, Edmund W. "Elementary and Persistent Errors in the Economic Analysis of Intellectual Property" 53 *Vanderbilt Law Review*, November 2000, p. 1727.

Landes, W., and R. Posner. "Indefinitely Renewable Copyright." University of Chicago Law Review, 70(2), 2003, 471–518.

Li, Xing, Megan MacGarvie, and Petra Moser. "Dead poets' property—how does copyright influence price?" *The RAND Journal of Economics* 49, no. 1 (2018): 181-205.

Liebowitz, Stan J. "The Myth of Copyright Inefficiency" *Regulation Magazine*, Spring 2009.

Liebowitz, Stan J. "Paradise Lost or Fantasy Island? Voluntary Payments by American Publishers to Authors Not Protected by Copyright." *The Journal of Law and Economics* 59, no. 3 (2016a): 549-567.

Liebowitz, Stan. "The Case for Copyright." Geo. Mason L. Rev. 24 (2016b): 907.

Liebowitz, Stan J., and Alejandro Zentner "The Challenges of Using Ranks to Estimate Sales" *Journal of Economics and Management Strategy*, DOI: https://doi.org/10.1111/jems.12552

Electronic copy available at: https://ssrn.com/abstract=4547458

Mortimer JH. “Price discrimination, copyright law, and technological innovation: Evidence from the introduction of DVDs” *The Quarterly Journal of Economics*, 2007

Publishers Weekly. “BiblioBazaar: How a Company Produces 272,930 Books A Year” Andrew Albanese, April 15, 2010.

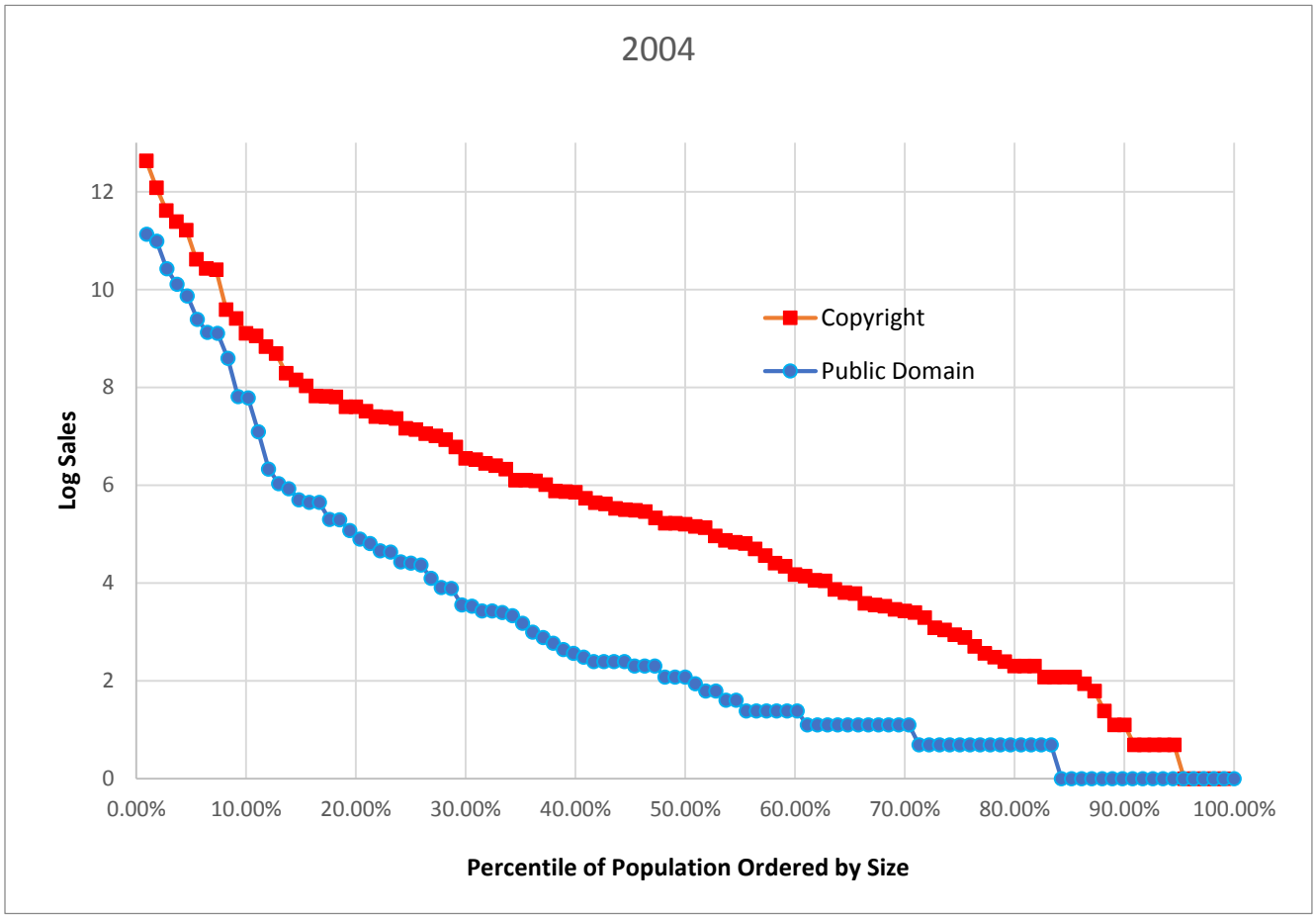
Publishers Weekly. “Self-Published Titles Topped 764,000 in 2009 as Traditional Output Dipped” Jim Milliot, Apr 14, 2010.

Reimers, Imke. "Copyright and generic entry in book publishing." *American Economic Journal: Microeconomics* 11, no. 3 (2019): 257-84.

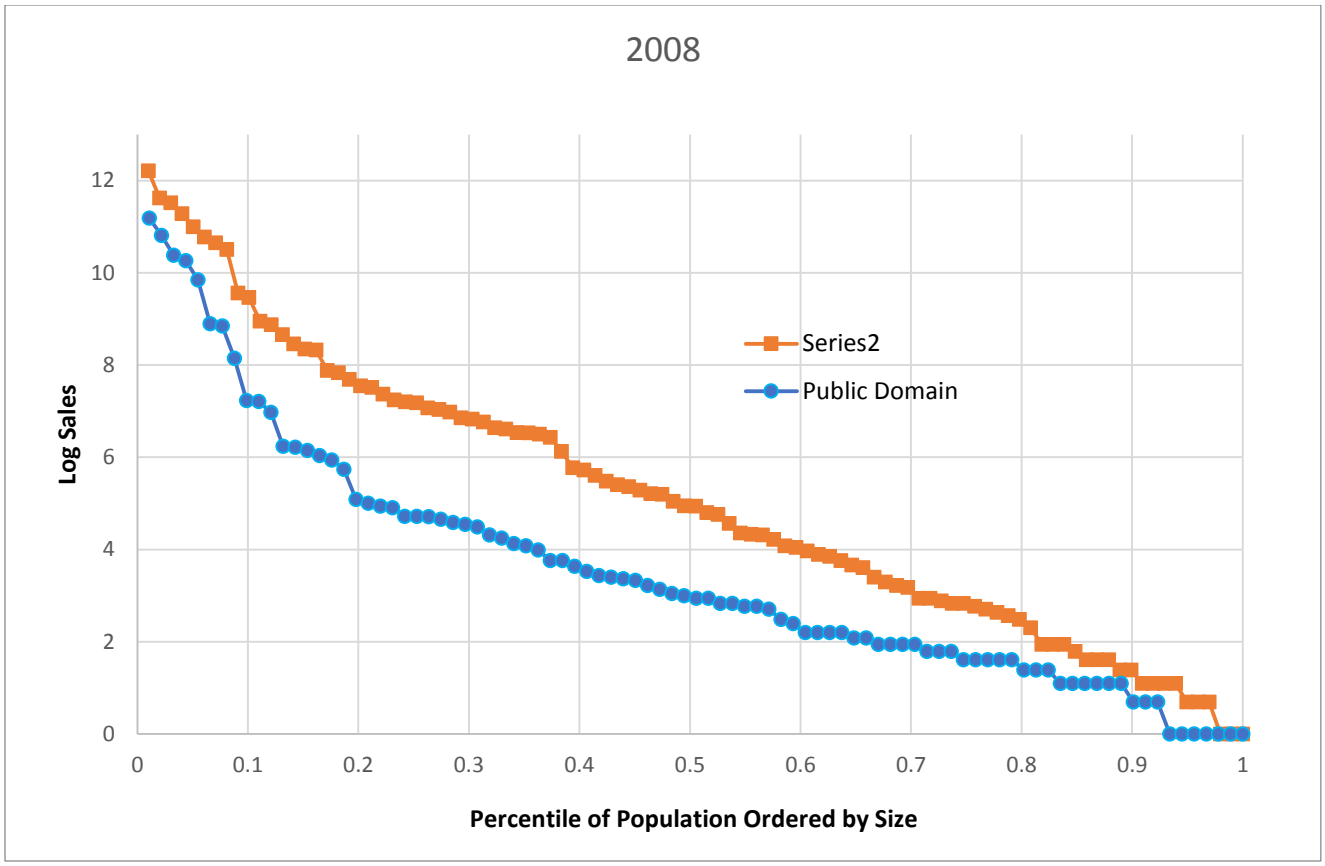
Electronic copy available at: https://ssrn.com/abstract=4547458

Appendix 1: CR to PD sales in multiple years

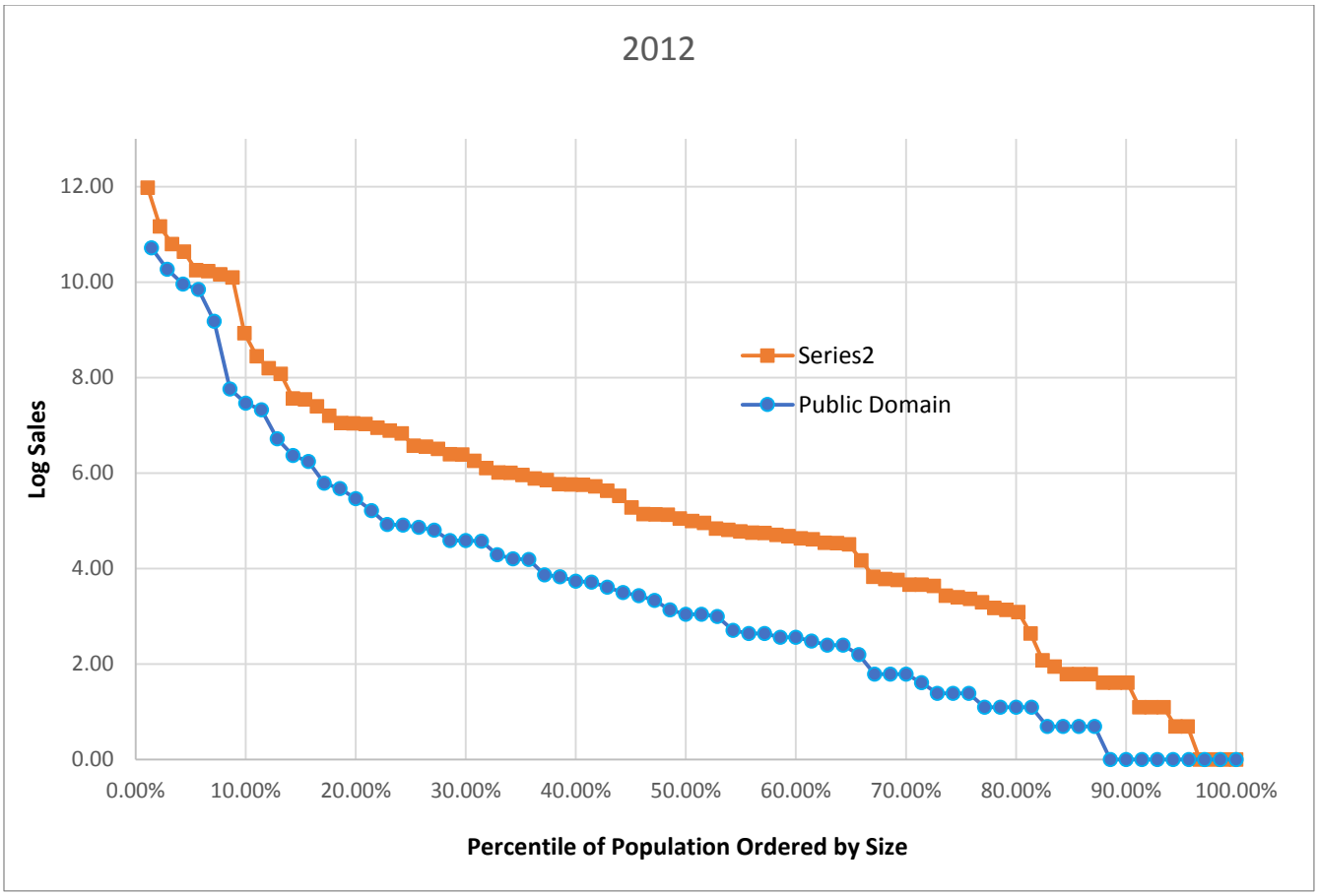
In the main text we show how CR sales are larger than PD sales in 2004, which is a key result. Here we show the equivalent diagrams to Figure 1 in the text, for all four years, 2004, 2008, 2012, and 2016. Although not identical, the figures are all similar in showing that CR sales (logged) are greater than PD sales (logged) for every similarly situated title. At the end we show what Figure 1 looks like if the sales values are not logged.



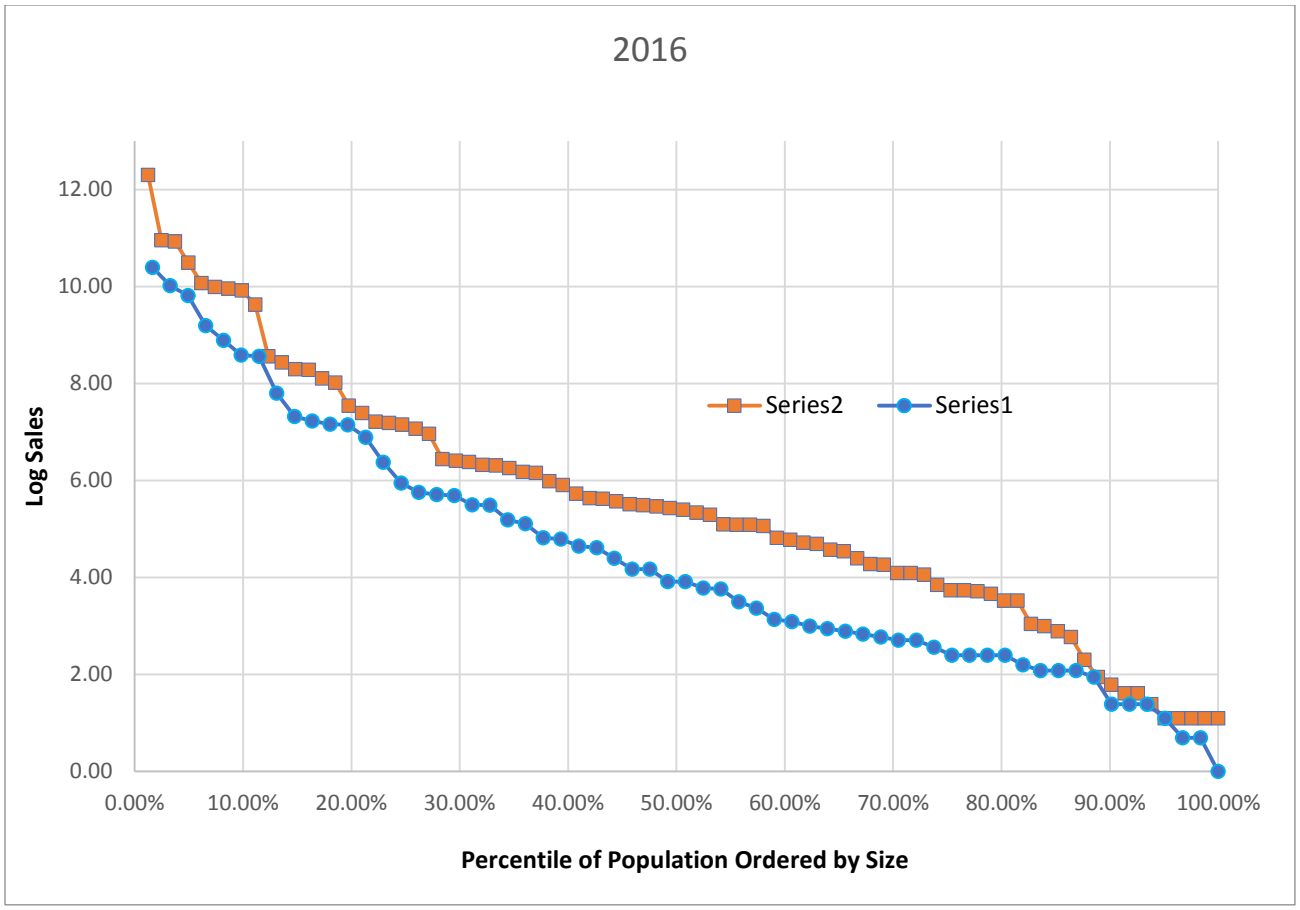
Electronic copy available at: https://ssrn.com/abstract=4547458



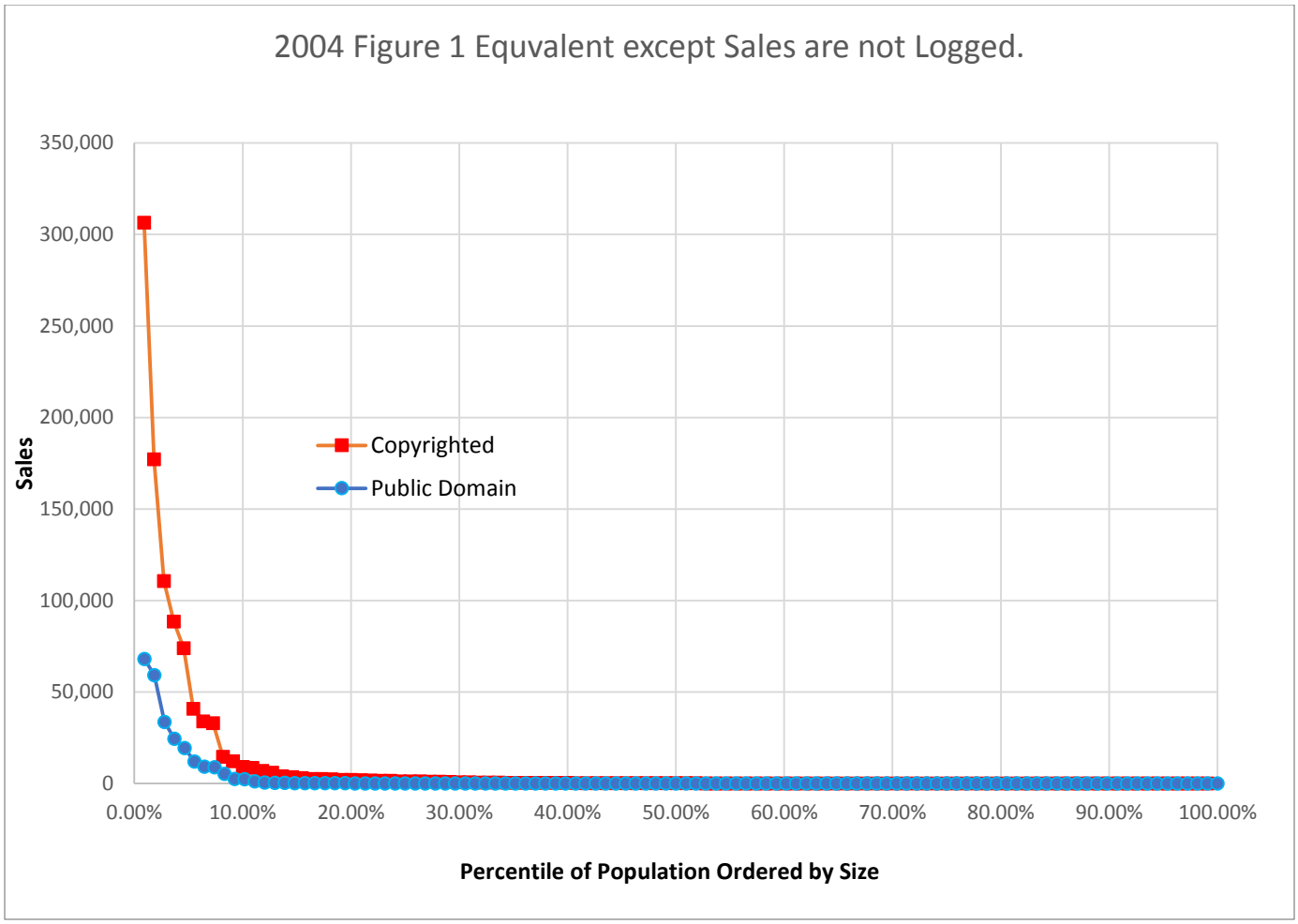
Electronic copy available at: https://ssrn.com/abstract=4547458



Electronic copy available at: https://ssrn.com/abstract=4547458



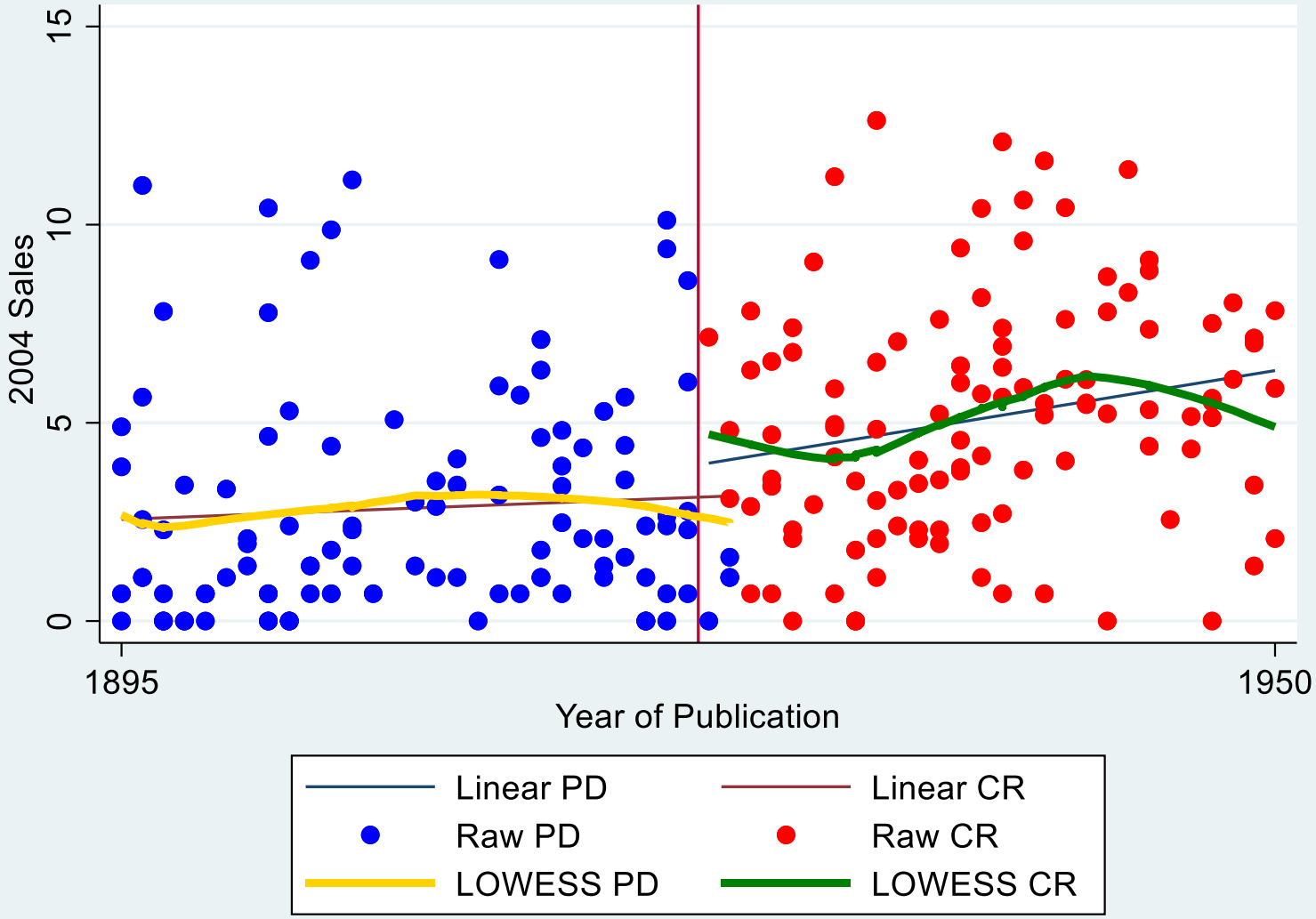
Electronic copy available at: https://ssrn.com/abstract=4547458



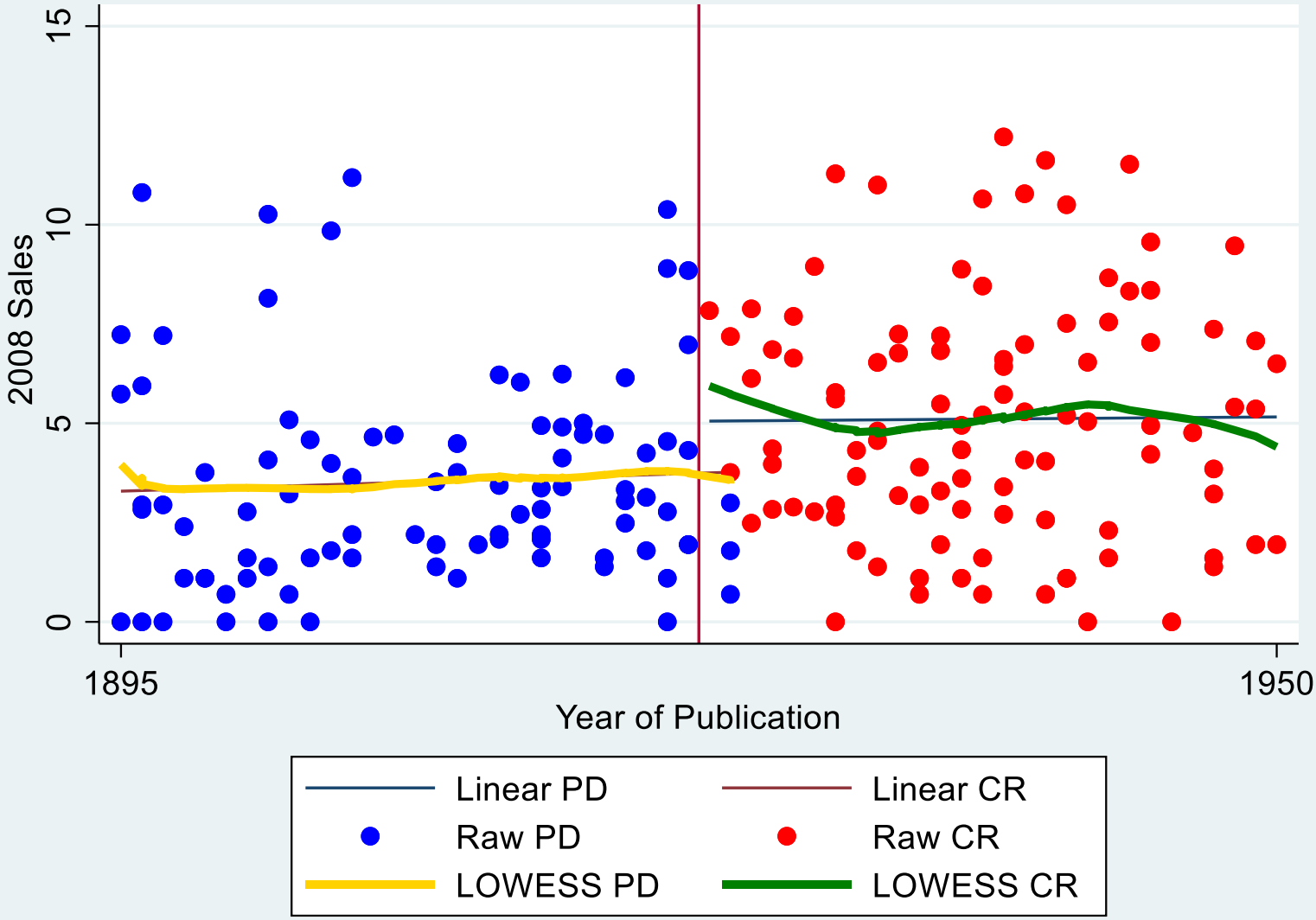
Electronic copy available at: https://ssrn.com/abstract=4547458

Appendix 2: Title Sales by Year Figures

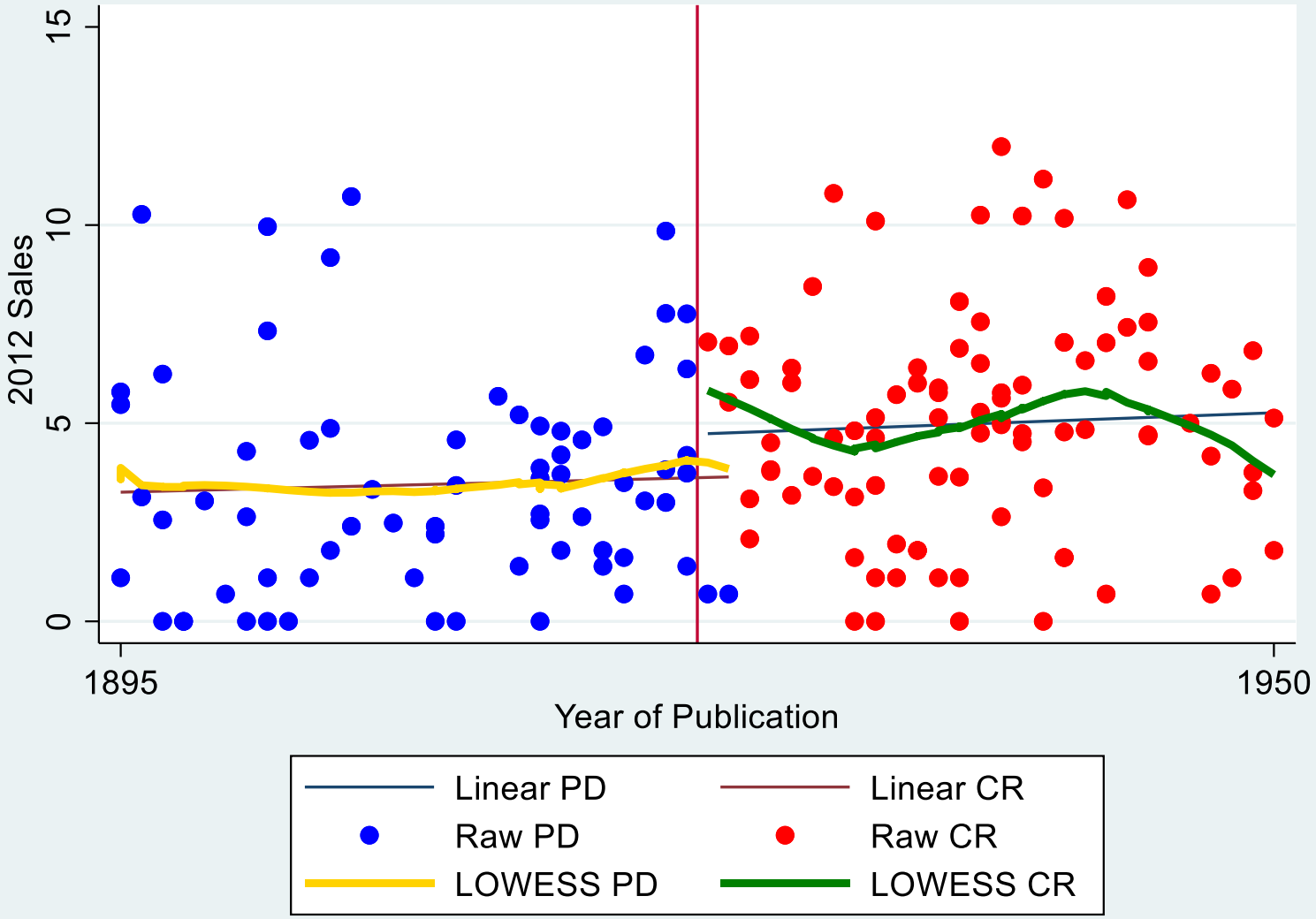
This Appendix provides yearly equivalent versions of Figures 2 and 3 in the paper for all four years 2004, 2008, 20112 and 2016. First we will show the figures for logged sales (Figure 2 equivalents) and then we will show the figures for unlogged sales values (Figure 3 equivalents).



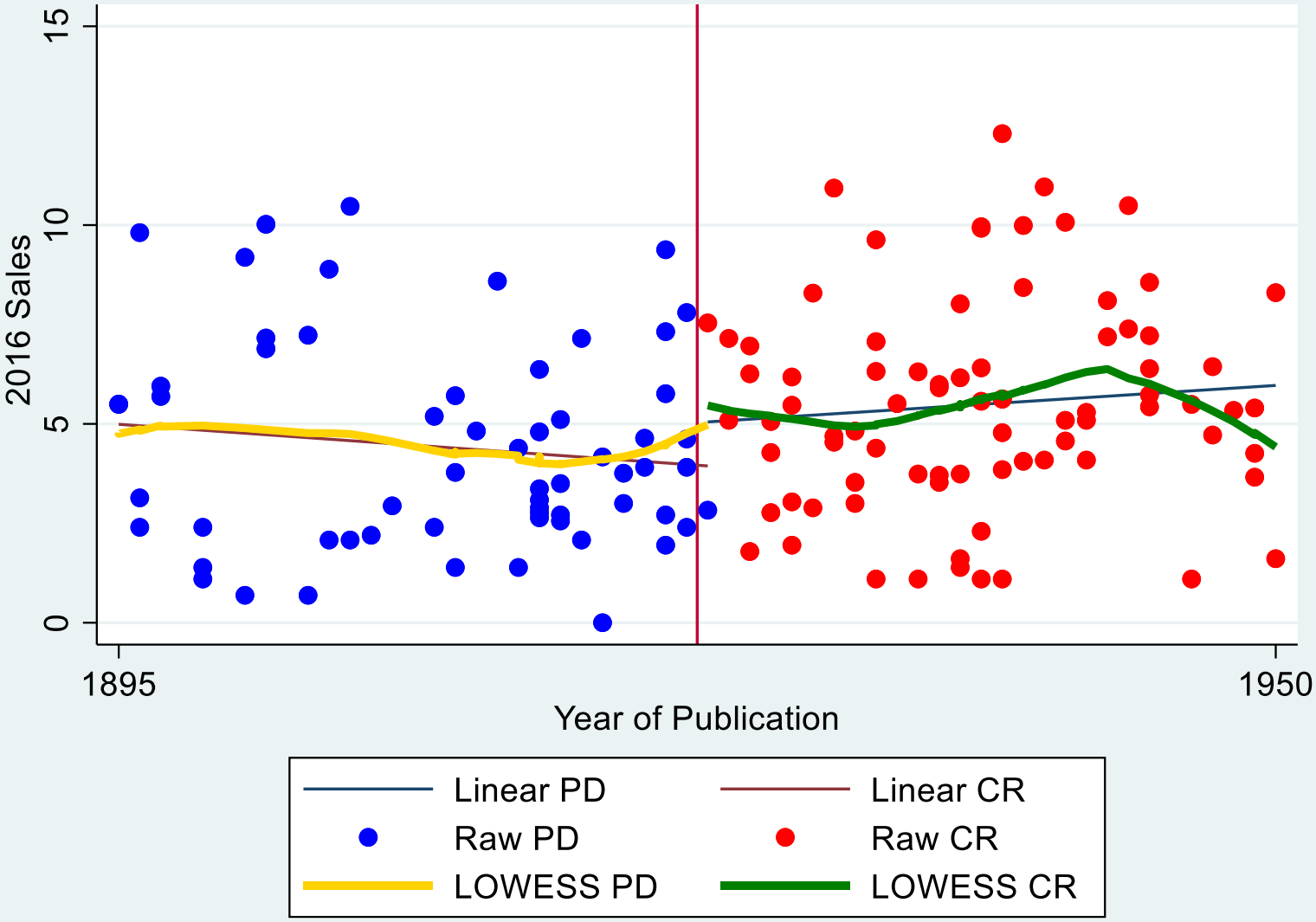
Electronic copy available at: https://ssrn.com/abstract=4547458



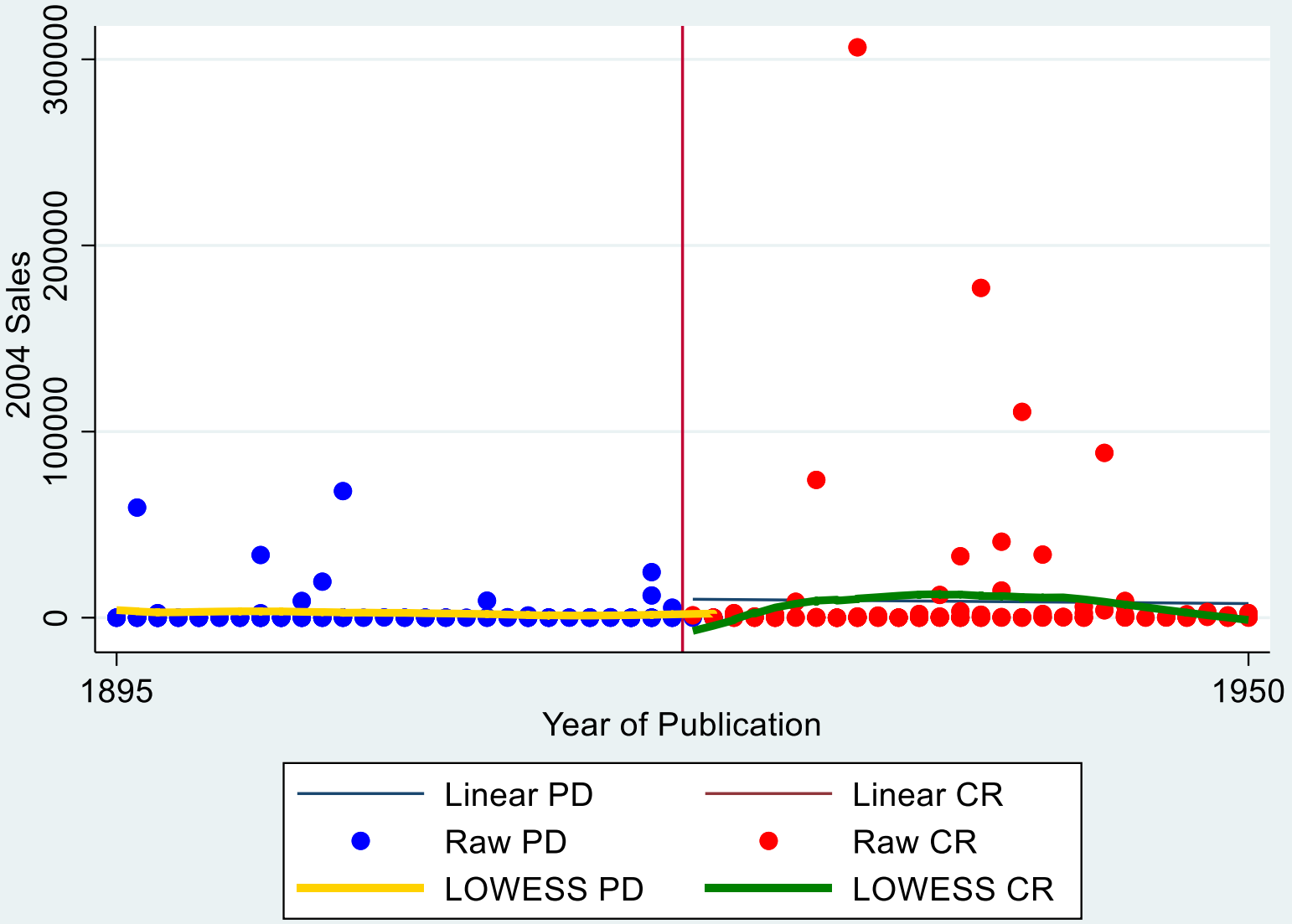
Electronic copy available at: https://ssrn.com/abstract=4547458



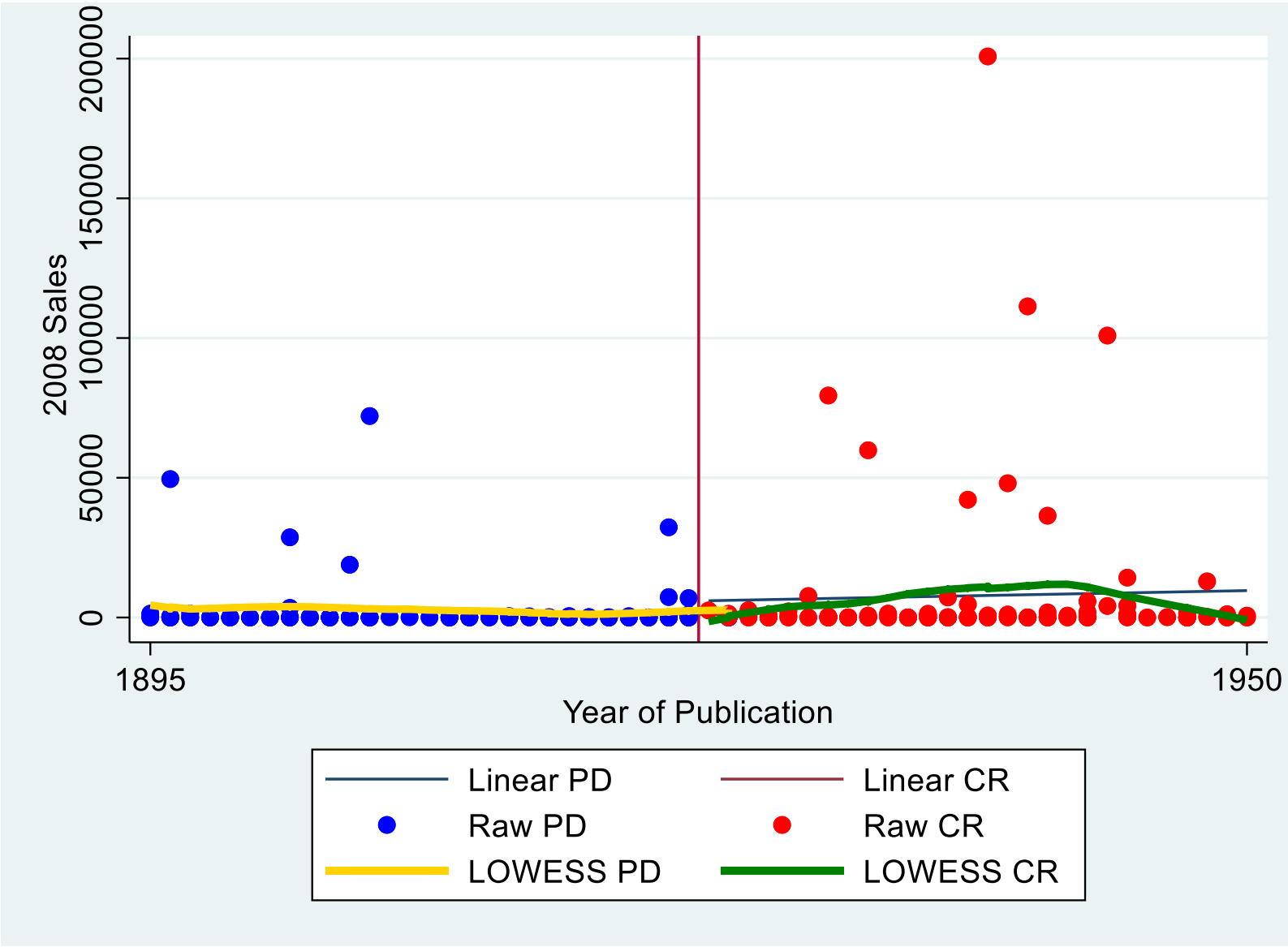
Electronic copy available at: https://ssrn.com/abstract=4547458



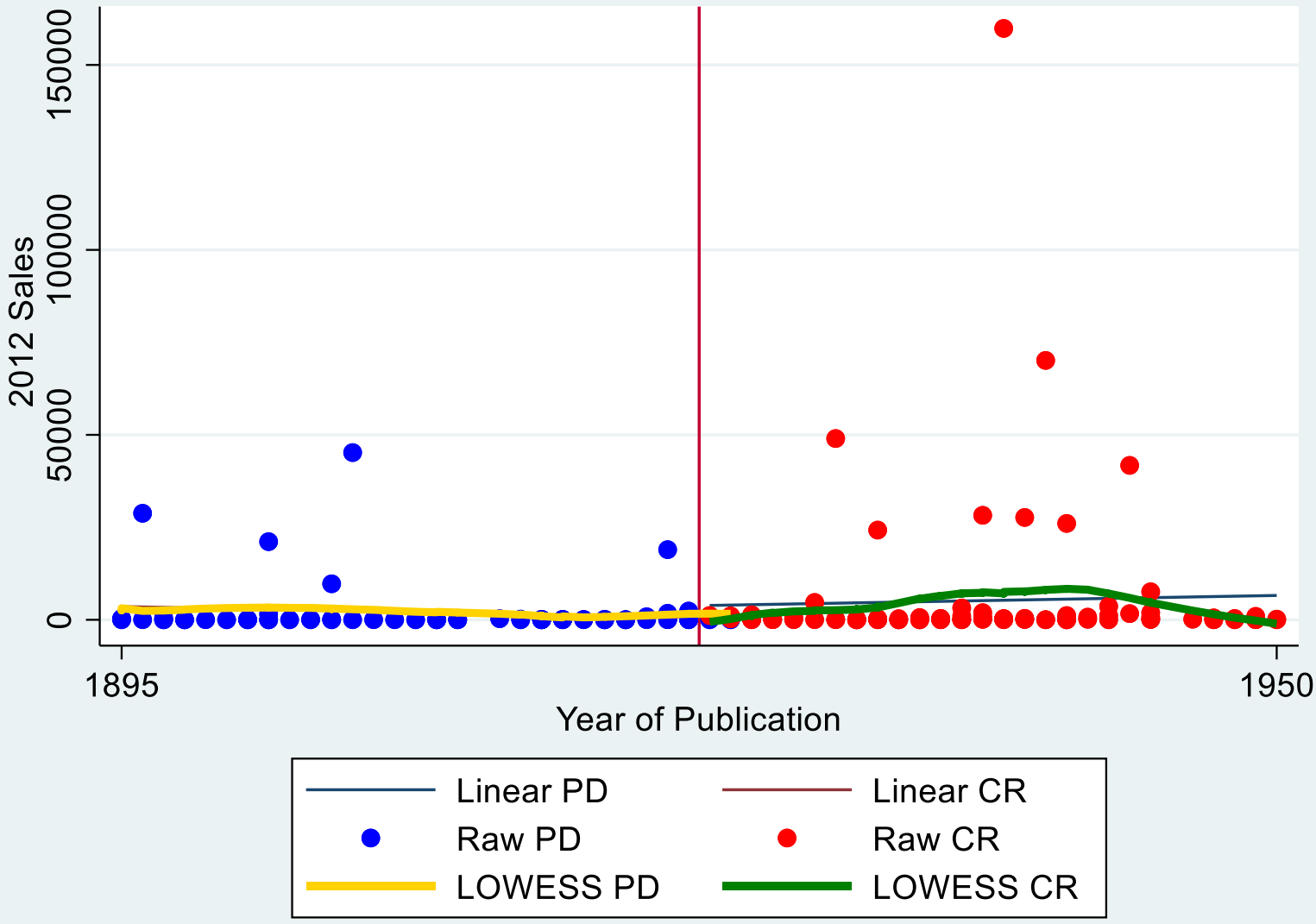
Electronic copy available at: https://ssrn.com/abstract=4547458



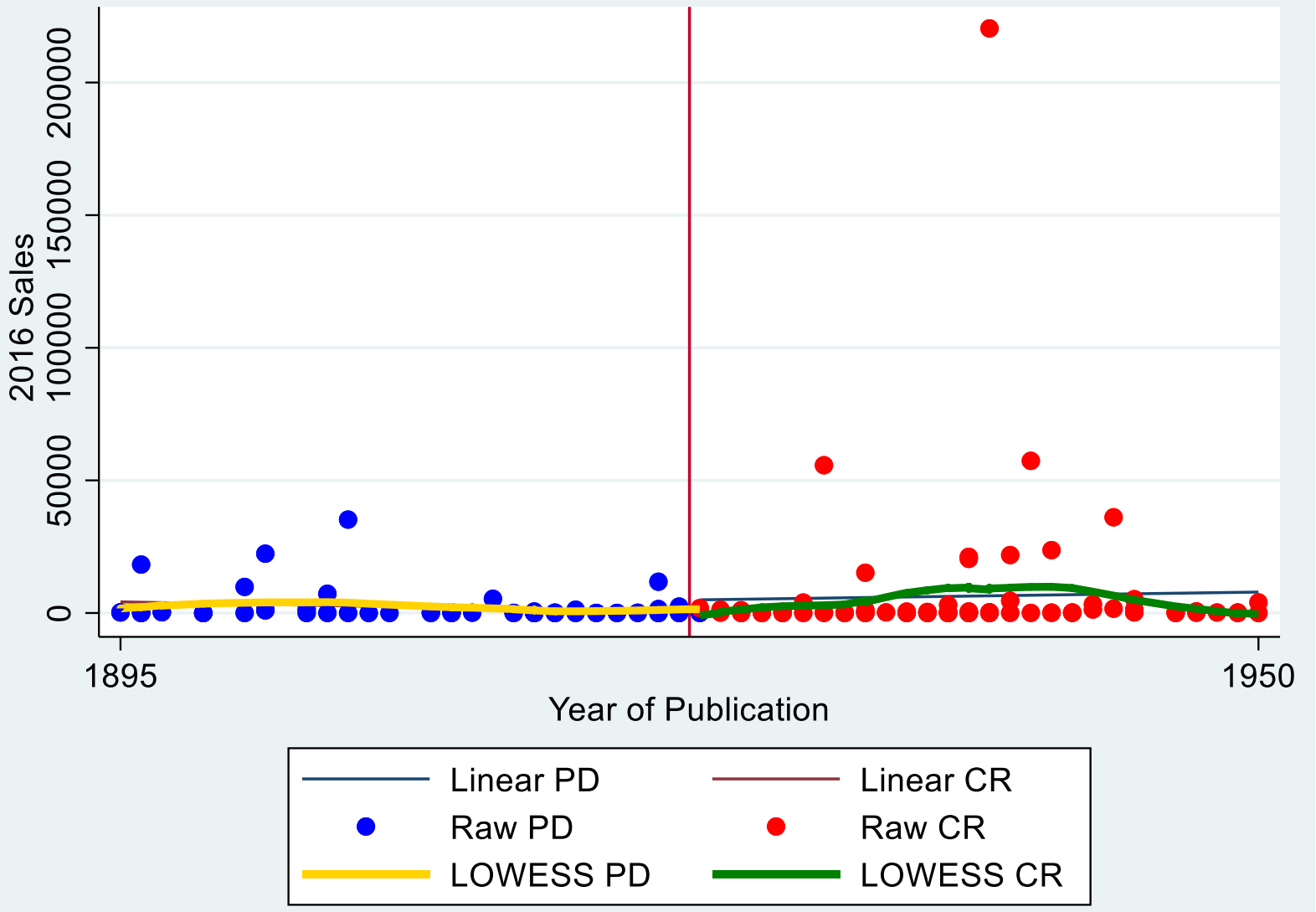
Electronic copy available at: https://ssrn.com/abstract=4547458



Electronic copy available at: https://ssrn.com/abstract=4547458



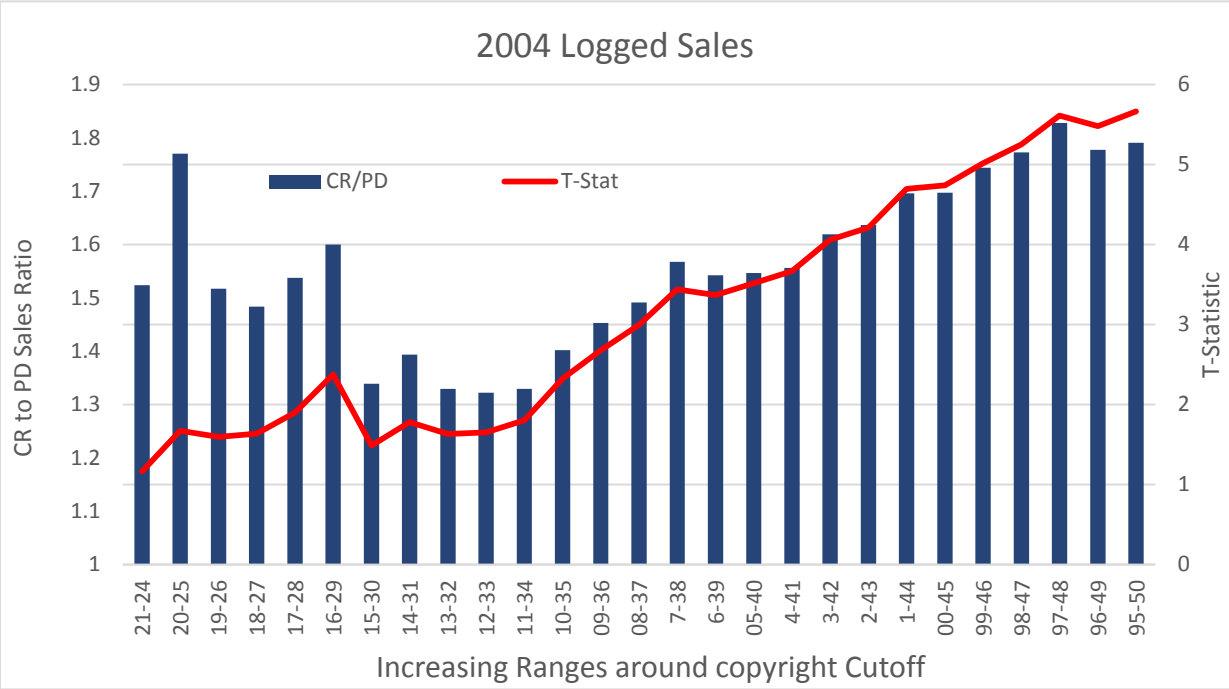
Electronic copy available at: https://ssrn.com/abstract=4547458

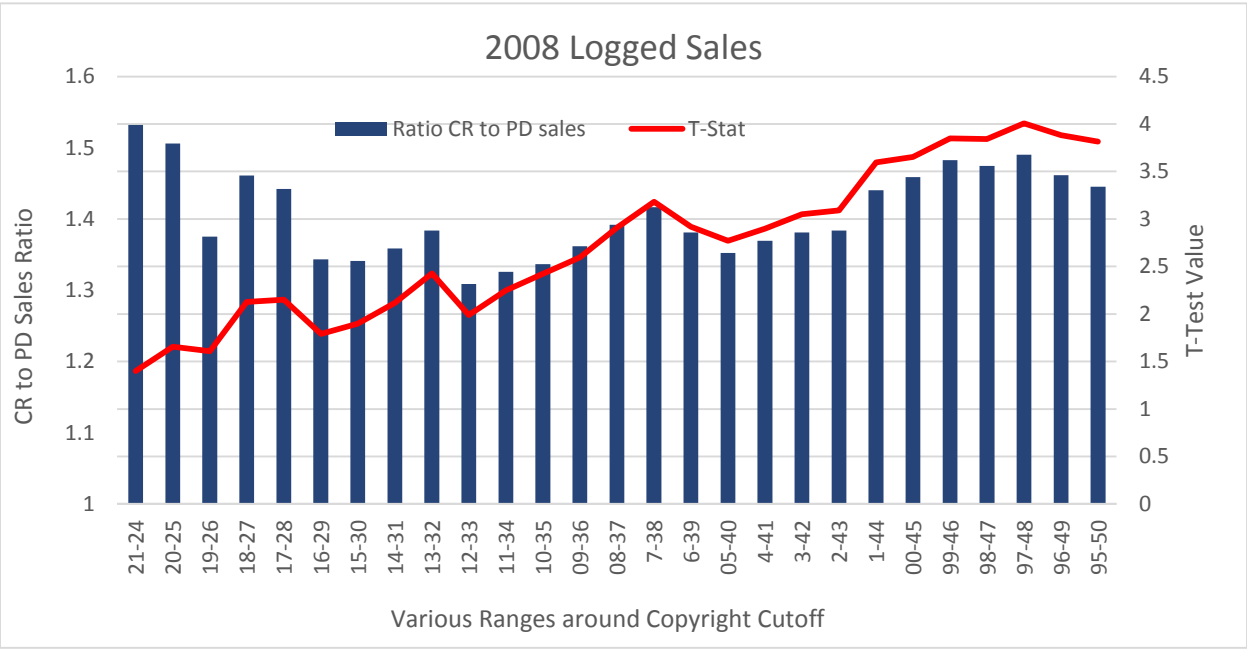


Electronic copy available at: https://ssrn.com/abstract=4547458

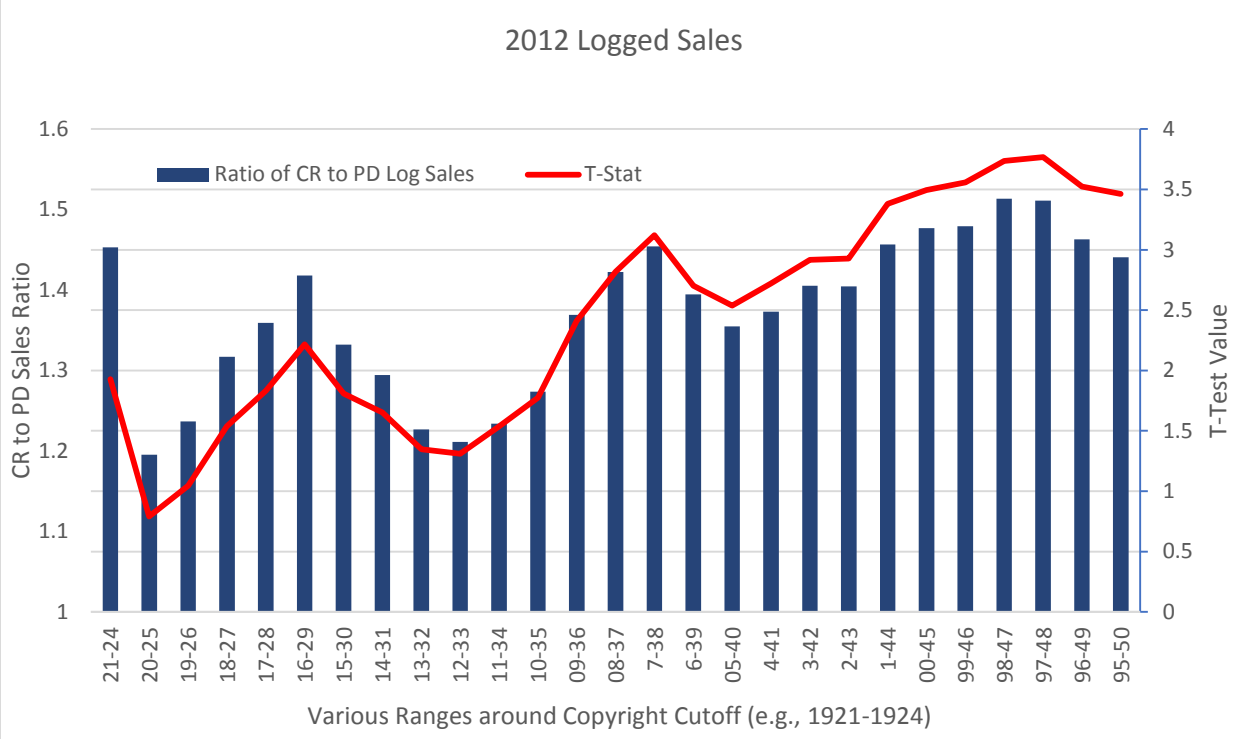
Appendix 3: CR to PD Ratios for Various Year Ranges

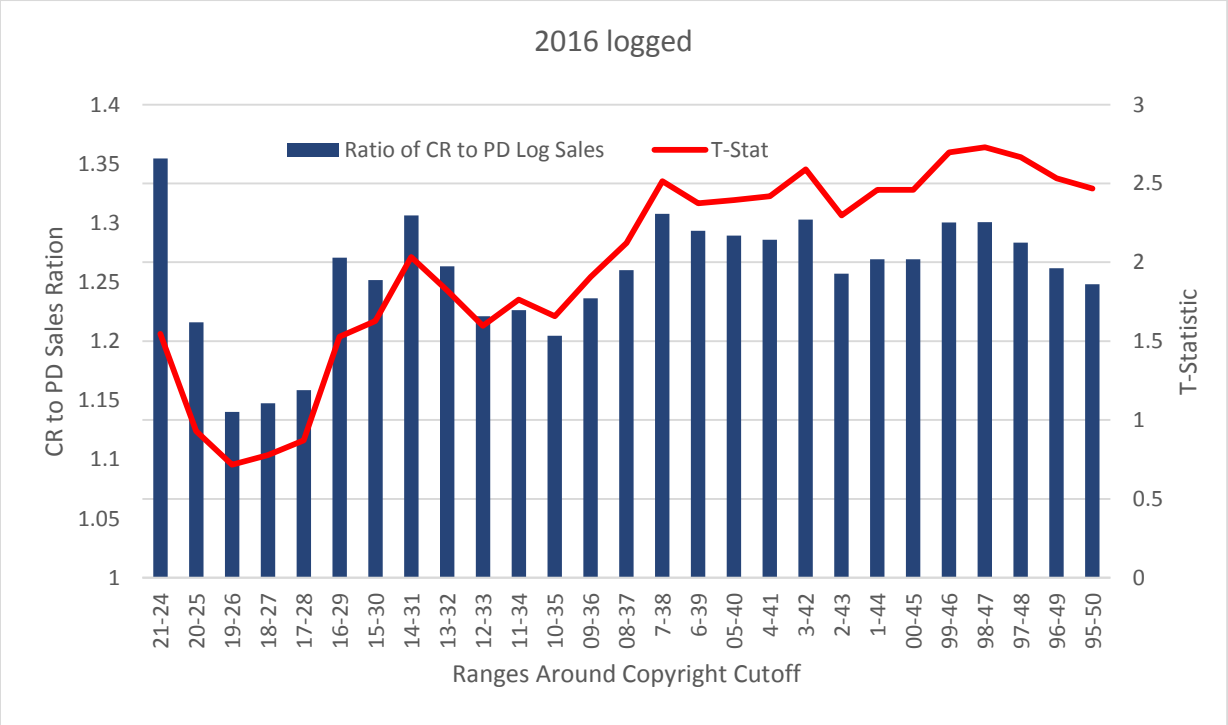
In the main text we show how the CR to PD sales ratio changes for different year ranges around the 1922/3 copyright cutoff. Here we show the diagram, Fig 4 in the text, for all four years, 2004, 2008, 2012, and 2016. Although not identical, the figures are all similar in showing that CR sales (logged) are greater than PD sales (logged). We also show these figures with unlogged values for the same four years. Note that the values on both vertical axes change from diagram to diagram.





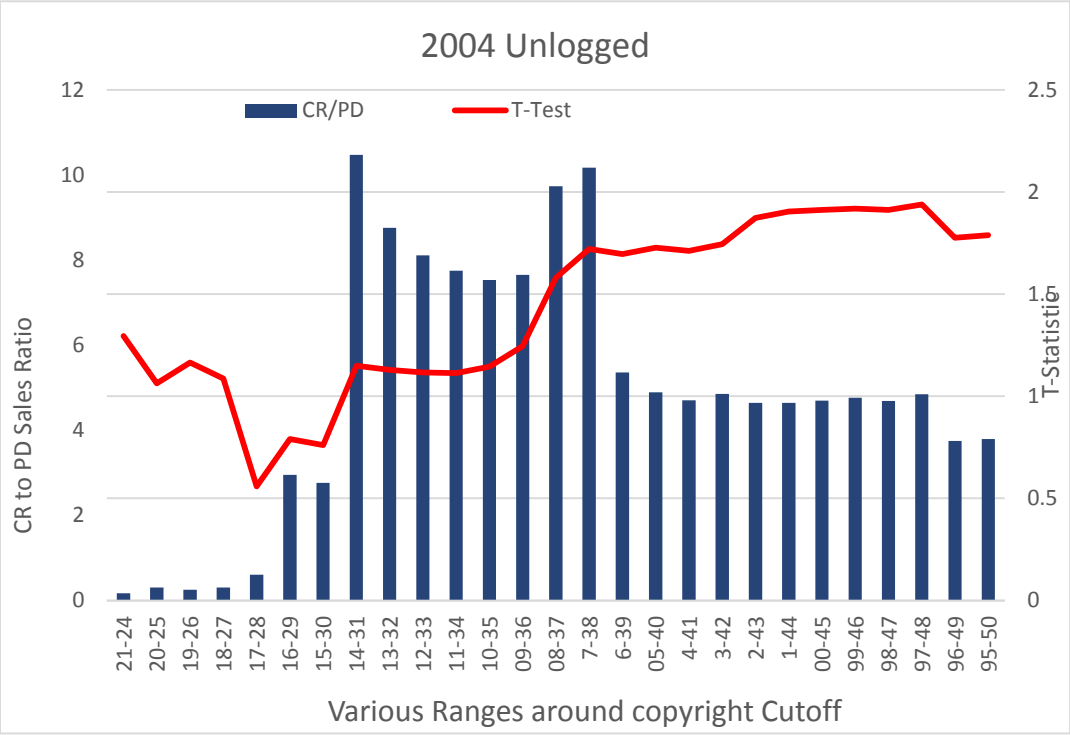
Electronic copy available at: https://ssrn.com/abstract=4547458



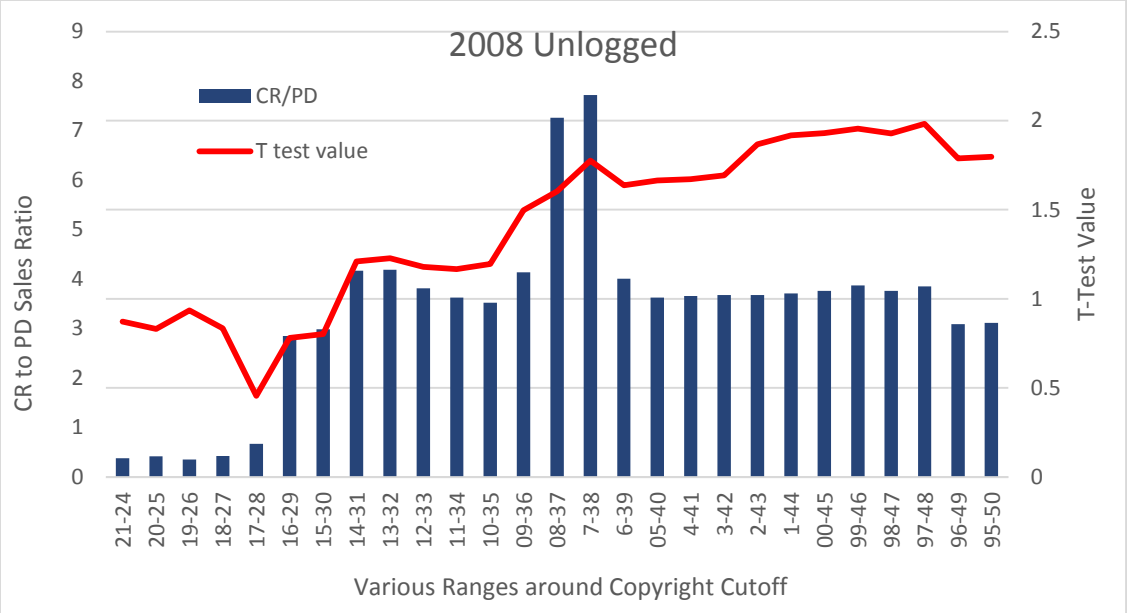


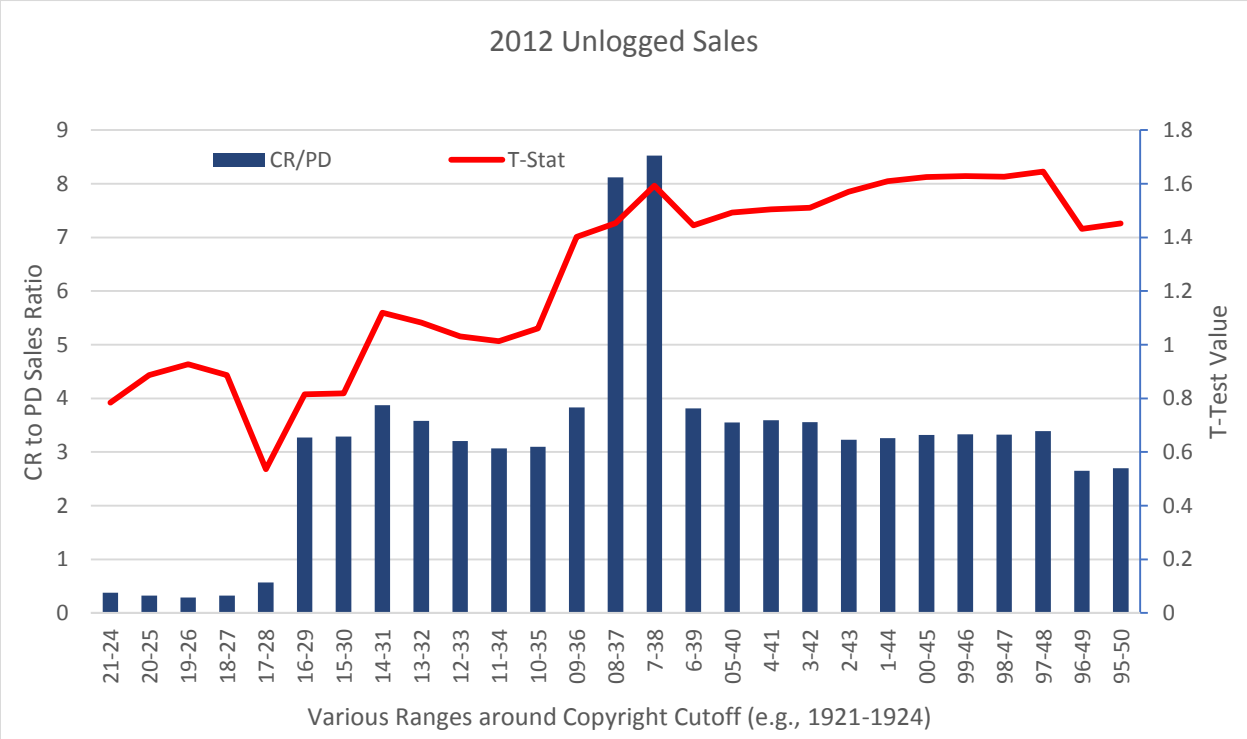
Electronic copy available at: https://ssrn.com/abstract=4547458

The key difference between logged and unlogged sales is that the first five cutoff possibilities have PD sales larger than CR sales. That is because there are no large (outlying unlogged CR titles until the 6th year whereas there are some PD titles with such values. This can be seen in the main paper if you look at Figure 3. Going left from the vertical line in that figure there are 3 PD outliers (define as a visibly separate circle) in the first two years going left and none going right until the 5th and 6th years. That is the danger of using the highly skewed unlogged data.

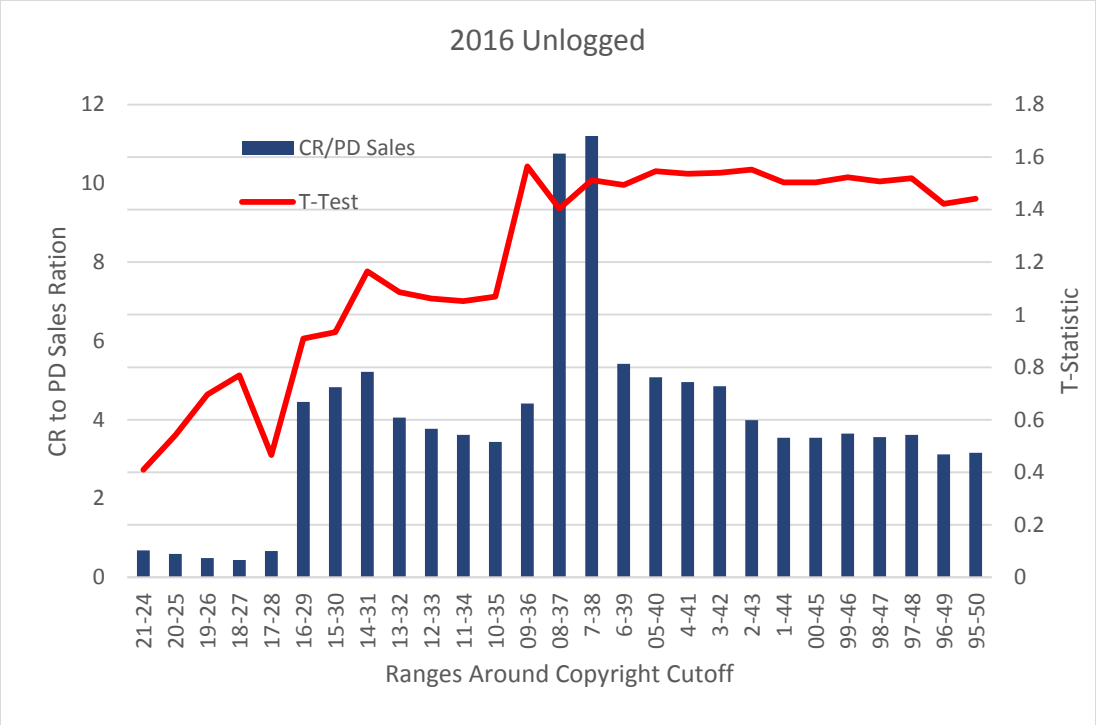


Electronic copy available at: https://ssrn.com/abstract=4547458





Electronic copy available at: https://ssrn.com/abstract=4547458



Electronic copy available at: https://ssrn.com/abstract=4547458

Appendix 4: Economies of Scale

This appendix represents a preliminary sketch of book industry prices, sales, and costs and does not claim to be in any way authoritative.

A priori, virtually everything we know about printing expenses indicates that there are economies of scale in book production. With the enormous range of our data, we can be certain that the top selling titles have much larger demands than the poorer selling titles. We normally associate larger demands with higher prices, but that is because we believe supply curves are upward sloping. If economies of scale are large enough, supply curves might be downward sloping.

Our data set does not have production costs and cannot, therefore, directly test for economies of scale. Nevertheless, there are reasons to believe that market prices would reflect economies of scale. First, Clerides presents evidence that (hardcover and paperback) book prices reflect cost shifters but not demand shifters which would imply the economies of scale would show up in prices. Second, if markets are competitive, we would expect to see lower costs reflected in lower prices. Third, even if markets are not competitive, we would also expect to see lower costs reflected in lower prices if the degree of monopoly power were unchanged as output changed.

If market prices reflected economies of scale, there should be a negative relationship between sales levels and price, holding other cost shifter constant. When you first look at the unconditional BookScan data there appears to be support for this supposition. Table 41 contains coefficients and t-statistics from simple regressions of price on log sales. There are four years of data and in each year there are four samples of data examined, so there are 16 coefficients representing results from 16 regressions. The first row of coefficients, based on all formats of BookScan fiction editions combined, suggests a negative relationship between sales and price, although it is considerably weaker in 2016 (but as noted in the main text, data in 2016 are less reliable than other years due to missing observations). These coefficients also do not control for other factors such as formats or number of pages, making these results incomplete.

Table 41: Coefficient of Price on Log Sales, BookScan Fiction

	2004	2008	2012	2016
Coef (full 126k data)	-0.77332	-0.61058	-0.64847	-0.05299
t-stat	-98.04	-47.00	-67.42	-5.43
Coef (Hardcover)	-0.19642	-0.42462	-0.69799	0.325091
t-stat	-12.68	-9.43	-24.62	8.28
Coef (Trade Paper)	-0.27002	-0.40573	-0.18615	0.142792
t-stat	-32.7	-55.48	-22.48	17.88
Coef (MM Paper)	0.227696	0.285622	0.274457	0.234
t-stat	66.86	129.81	93.8	65.88

Electronic copy available at: https://ssrn.com/abstract=4547458

When one attempts to control for the most important cost shifter, format, as is done in the next three rows of Table 41, however, the simple negative relationship changes. The three sets of rows after the first compare the relationship between sales and price for each of the three major formats. Hardcover and Trade Paperbacks have a negative relationship, except in 2016. Strikingly, however, MM Paperbacks have a consistent positive significant relationship between sales and price in each year.

It should be noted that MM Paperbacks are quite different than the other two book categories. Instead of being sold in online or offline bookstores, MM Paperbacks tend to be sold in supermarkets, drugstores, and airport retailers. Also, MM Paperbacks are usually from well-known authors and have previously sold well as hardcover or trade paperback editions. MM Paperbacks are almost exclusively published by major publishers. Is there any reason to believe that MM Paperbacks will not exhibit economies of scale in production? The answer is "no." MM sales are considerably larger on average than are those of hardcover or trade paper editions (about 3:1) so they are more likely to be printed using offset printing, which exhibits strong economies of scale.

That leaves the question of how to interpret the strong positive coefficient for MM editions? It is possible that, given the low prices and costs of MM editions, that the absolute average cost reductions are small as output increases and that other variable costs that increase on average with output overwhelm the small savings in production costs for MM editions. But without further evidence on this point, this is merely speculation.

We can investigate whether this relationship also holds for our vintage titles. Table 42 examines the relationship between price and logsales for vintage titles in 2004. Row 1 shows a large negative coefficient between price and logged sales when we don't distinguish between formats, as was the case in Table 41, except that the coefficient is even larger. The next three rows also provide results similar to that of Table 41, where trade paper and hardcover editions have significant negative relationships between sales and price (also larger than Table 41), whereas MM Paperbacks do not have such a relationship, although the positive relationship for MM is weaker than that found in Table 41. Our sample of several hundred vintage titles shows the same general relationships between logged sales and prices as our much larger BookScan database of several hundred thousand observations.

Table 42: Coefficient of logprice on sales, Vintage Editions 2004

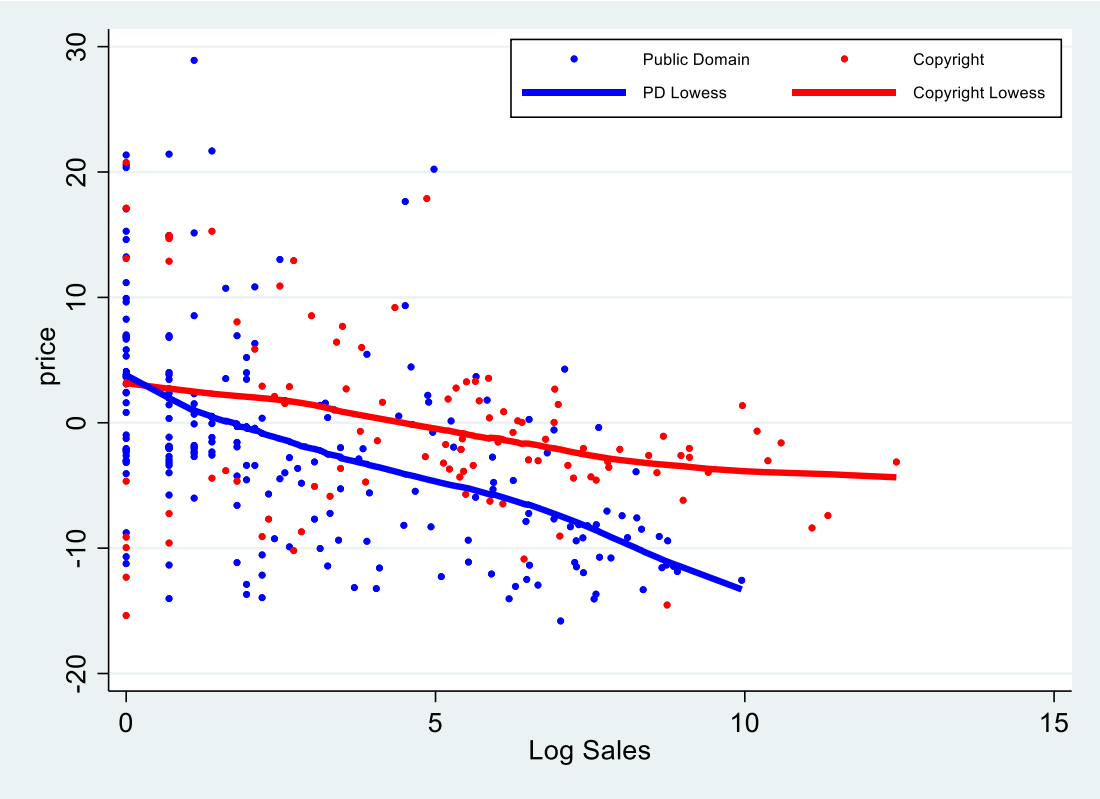
	coefficient	t-stat	observations
Full Vintage	-1.48	-11.62	447
Vintage Trade Paper	-1.06	-6.88	260
Vintage Hardcover	-0.96	-2.68	124
Vintage MM	0.08	0.9	62

Electronic copy available at: https://ssrn.com/abstract=4547458

Our data on vintage editions, with about half the titles being in the public domain as opposed to the entire industry where almost all the editions are copyrighted, allows us to examine whether copyrighted titles behave differently with respect to prices and sales than public domain titles.

Figure 41 shows the relationship between sales and price (reflected in the LOWESS curves) for trade paperback editions, controlling for pages and separated by whether titles are CR or PD.[1] Because we are controlling for pages, the prices on the vertical axes are the residuals, which is why there are negative values.

Figure 41: Sales and Price for Vintage Trade Paper Editions by Copyright Status controlling for pages, 2004



The LOWESS smoothing lines indicate that the prices of both CR editions and PD editions tend to fall as sales increase. They also indicate that CR titles of a given sales level tend to have higher prices than PD titles with equivalent sales, and the difference seems to get larger as sales get

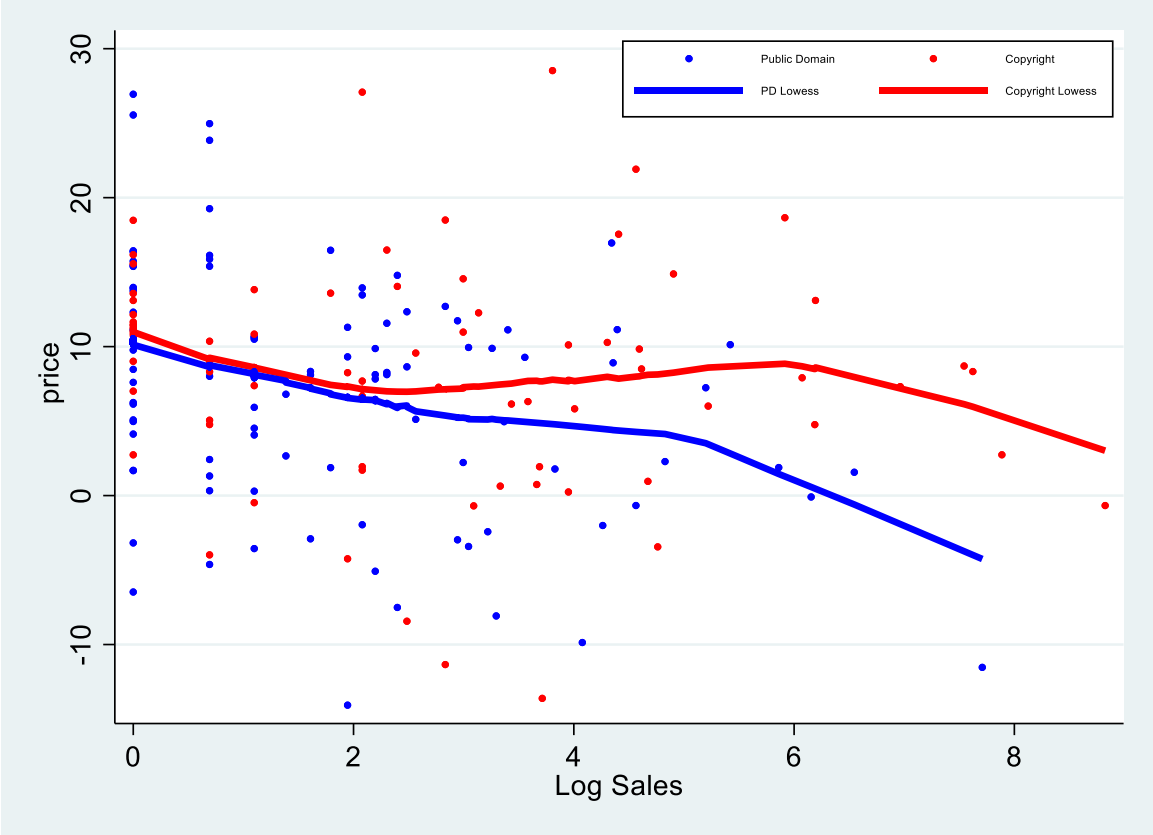
[1] The control for pages is based on the predicted impact of pages on price for the entire sample, not just the sample upon which the figures are based. This seems preferable because we know that increased pages increase cost yet if we run a regression for mm paperbacks only, the coefficient on pages is statistically an insignificant with a negative sign. In most cases the diagrams are very similar whether we control for pages or not.

Electronic copy available at: https://ssrn.com/abstract=4547458

larger.[2] The steeper decline in price for PD editions can be taken to imply that publishers of PD editions pass on more of the cost savings from economies of scale than do publishers of CR editions. The fact that publishers of CR titles must pay copyright owners a royalty would not explain the different slopes, although it would explain the higher CR line. Instead, if there were an increasing copyright payment as sales increased, it could explain the different slopes. Although it is true that in the current market royalty payment rates frequently increase with sales, those escalators are based on lifetime sales of an edition and its binding types. Because all of our titles are many decades old is seems unlikely that the royalty rates would increase as sales increase. On the other hand, if post-creation investment was related to sales, those expenses might also be able to explain the different slopes of the two lines.

The LOWESS smoothing line for hardcover books is similar but with less of a downward slope, especially for copyrighted editions.

Figure 42: Hardcover books, 2004, control for pages



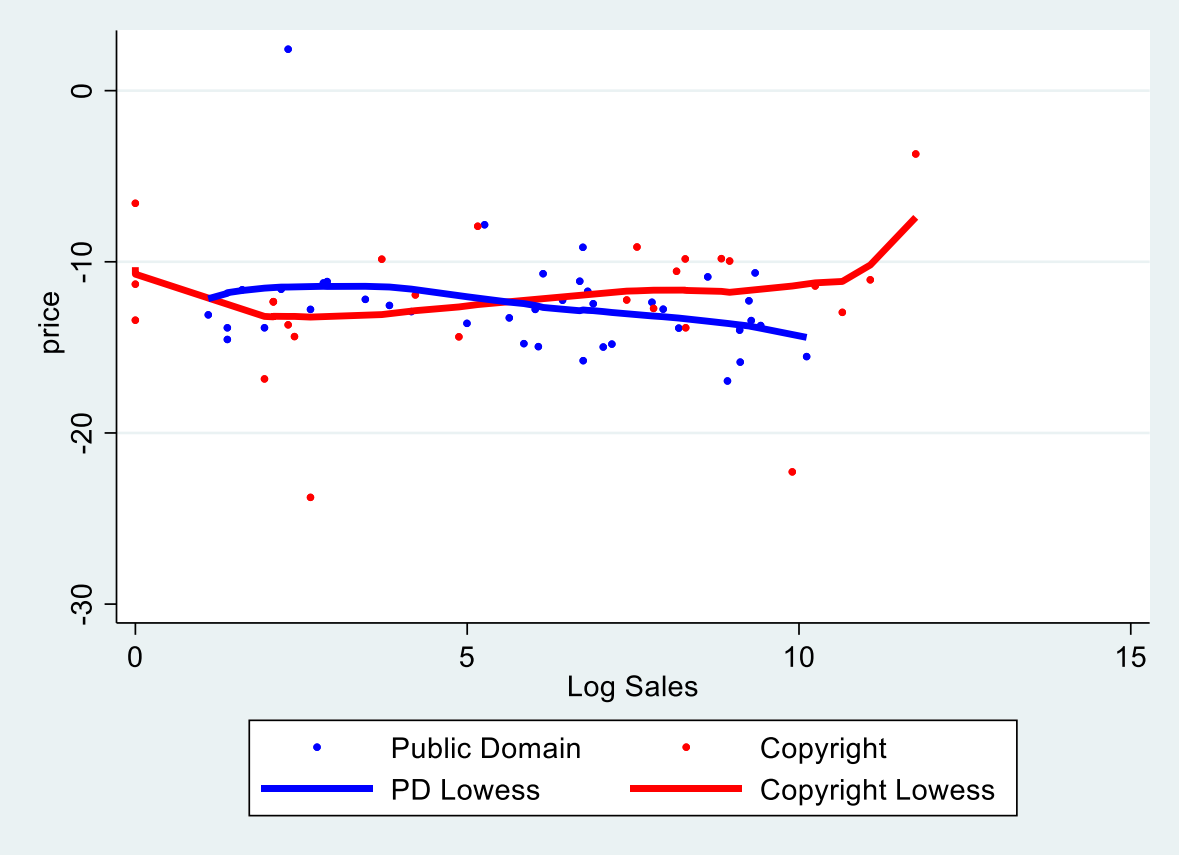
And for MM paperbacks, shown in Figure 43, the CR editions do not show a downward slope at all although the PD editions do show a slight decline over their range. For a portion of the range

[2] Consistent with the lowess curves, linear regression reveals an absolutely larger slope coefficient for PD editions (-141 versus -0.82).

Electronic copy available at: https://ssrn.com/abstract=4547458

PD editions have higher prices than CR editions, a relationship not seen in the other figures. Care needs to be taken before attributing too much to these figures as the number of observations is not large. As we saw before MM editions seem to have different economic characteristics than the other formats..

Figure 43: MM, 2004, control for pages



Because it is publishers who set the list price of an edition, and publishers who negotiate royalty rates and post-creation investment decisions, we might want to see whether major publishers and minor publishers have different behaviors.

Table 43: Vintage Paperback Sales by CR and Publisher Status, 2004

Sales	average	median
Minor CR	337	12
Minor PD	335	5
Major CR	28,241	6,271
Major PD	1,768	956

Electronic copy available at: https://ssrn.com/abstract=4547458

Table 43 reveals that for vintage paperback editions (which were seen in the main paper's Table 1 to contribute a majority of sales and editions), those editions controlled by major publishers sold many more copies than editions from minor publishers. The number of copies of CR editions and PD editions is not much different for minor publishers, but for major publishers CR editions sell many more copies than PD editions. The differences are stark, with CR editions from major publishers selling more than 80 times as many copies as CR editions from minor publishers and the difference in median values is even larger.

Figure 44, based on paperback editions, shows the relationship between sales and price for minor publishers (defined as any publisher not listed as a major publisher in the main paper's footnote 7). It looks like Figure 41, which represents the relationship for trade paper editions for both major and minor publishers. Note, however, that the horizontal axis ends at 10, which is a lower number than if you had included major publishers.

Figure 44 Sales and Prices for Minor Publishers of Trade Paperbacks

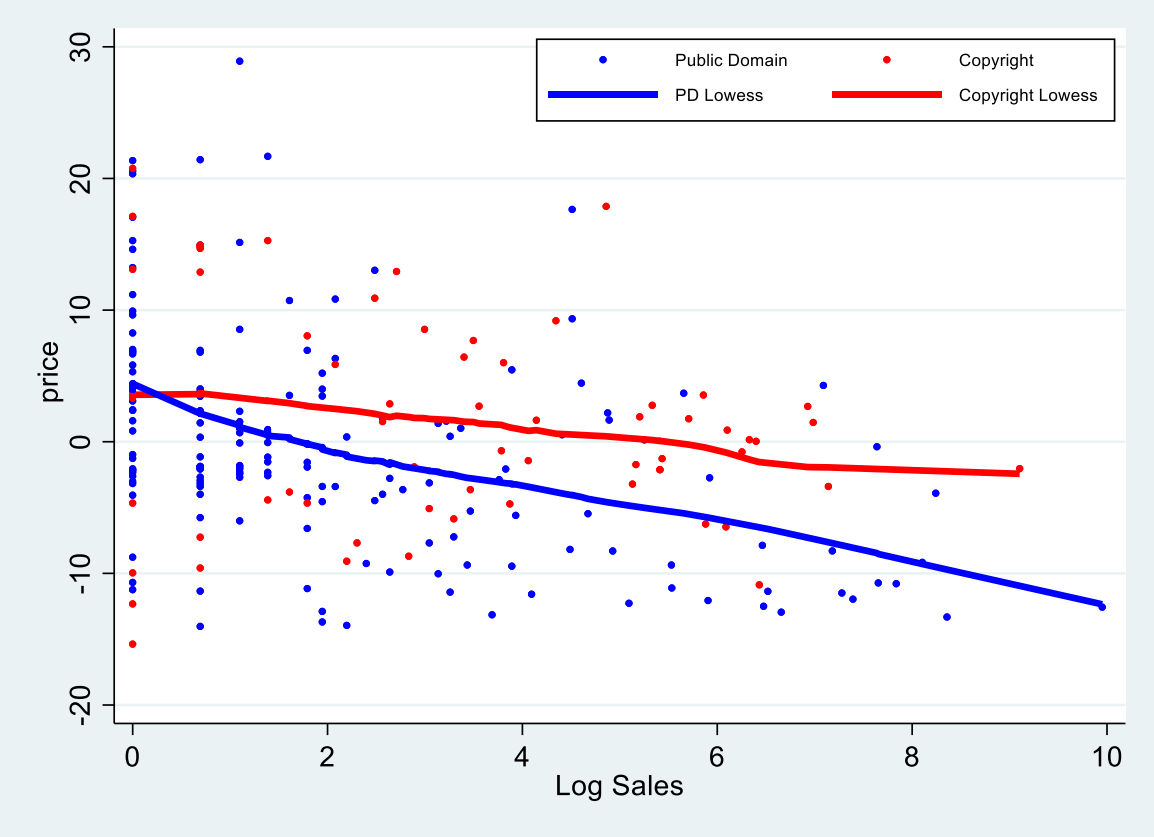


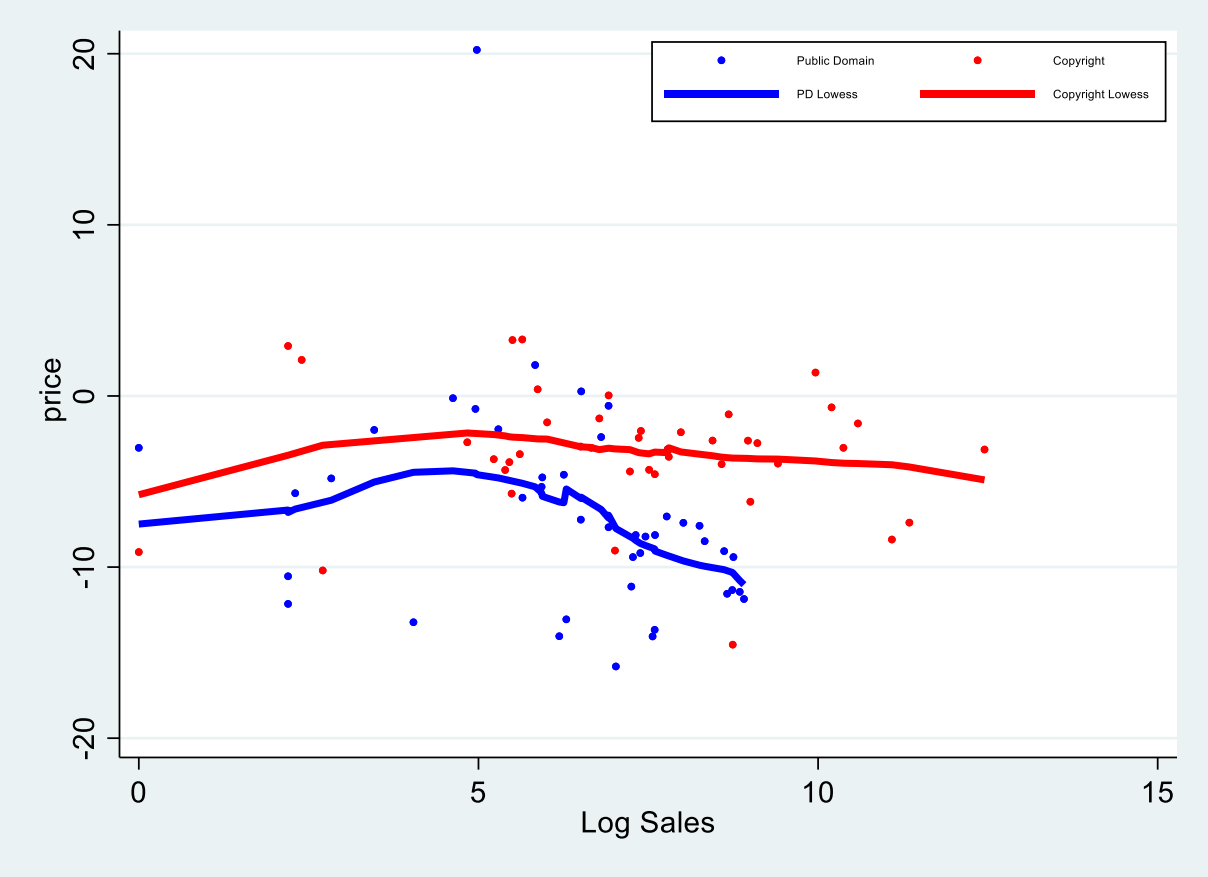
Figure 45 shows the same diagram but for major publishers. The CR LOWESS curve rises slightly and then falls slightly but is fairly flat overall. The PD LOWESS curve rises at first (due in large part to one unusually expensive edition) but then falls a fair amount for the better selling editions.

Electronic copy available at: https://ssrn.com/abstract=4547458

The better selling PD editions do not sell particularly well compared to the better selling CR editions. Note that the seven largest selling editions are all CR editions.

The CR editions from major publishers do not show signs of economies of scale although the PD editions do. Although we can talk of other variable costs overwhelming economies for MM titles, this is harder to do for trade paper editions since they tend to show economies of scale. One explanation might be that it is major publishers and their CR editions that are most likely to receive post-creation investment and economies of scale do not show up because of the expenditure on post-creation investment.

Figure 45: Trade Paper Sales and Prices, Major Publishers, 2004



It should also be mentioned, however, that Figure 45 is based on only a small number of observations (only 21 CR editions and 38 PD editions) so while it is intriguing, we cannot put too much weight on it. Also note that there are very few copyrighted editions on the upward sloping portion of the LOWESS curve, making it more likely to be misleading. The fairly horizontal (and slightly downward sloping) portion of the curve is based on more observations.

Electronic copy available at: https://ssrn.com/abstract=4547458

Still, the implication would be that like mm paperbacks, CR trade paper editions from major publishers do not seem to set price only on production cost shifters but seem to be affected by either post creation investment or demand.

Electronic copy available at: https://ssrn.com/abstract=4547458

Appendix 5: Discussion of Reimers

Reimers' data, although based on a set of vintage titles which is basically a subset of my vintage titles, is otherwise quite different. Her price data for editions are monthly and are based on retail values she scraped from Amazon's website. Her sales data are based on Amazon rankings that she scraped off the website, presumably every hour, and used some algorithm to convert improvements in rankings into sales quantities. Her data include both new and used editions of titles whereas my data do not include any information from the used book market.

Table 56: Reimers Table 3 (explaining log price) plus two changes.

Variable	As Published	Add Pages	Remove Used
Copyright	0.242***	0.203***	0.066
Standard Error (SE)	0.0757	0.0654	0.153
Pages		0.00110***	0.00117**
SE		0.000334	0.000405
eBook	-2.354***	-2.407***	-2.447***
SE	-0.142	-0.132	-0.128
Paperback	-0.656***	-0.611***	-0.595***
SE	-0.0407	-0.0432	-0.0458
Available New	-0.131*	-0.102	0
SE	-0.0623	-0.0687	(.)
Edition Age	-0.00379	-0.00388	-0.00632**
SE	-0.00262	-0.00276	-0.00275
Major Publisher	-0.194***	-0.244***	-0.362***
SE	-0.0644	-0.0591	-0.0973
N	2659	2659	2075
adj. R-sq	0.457	0.494	0.529

The first task in this appendix is to examine how copyright's impact on price is influenced by her choice of variables and her inclusion of used editions. Using her instructions for replicating her results, I was able to exactly reproduce her results. The first column of Table 51 reproduces her regression specification and results as found in her Table 3.[3] The number of observations is much larger than the number of titles she examines because there are 12 monthly observations for each edition. Based on the significant positive coefficient for copyright, found in the first row of

[3] The actual specification includes other variables such as monthly fixed effects, but I follow her Table 3 in only reporting details on the variables shown in the table.

Electronic copy available at: https://ssrn.com/abstract=4547458

Column 1, Reimers concludes that copyright significantly increases price by an amount in the mid-20% range (since the coefficient is only an approximation to the percentage value).[4]

In the second column I add her variable measuring the number of pages in an edition, which I consider to be fixing a simple oversight on Reimer's part. I don't think it is controversial to suggest that the number of pages in a book should be included in a regression explaining the price of the book. The inclusion of number of pages lowers the size of the (logged) copyright coefficient somewhat although it is still in the range of 20% and is still statistically significant.

In the third column I make a more substantial change to her method which changes the result considerably. As is obvious from the inclusion of her fifth variable, "Available as New," her observations include editions that are only available as a used book, or else this variable would have no meaning. If there is no new edition available, then she uses the used-book price for that edition. But we want to limit the analysis to new editions because copyright is irrelevant in the used book market. The first sale doctrine in the US says (Mortimer 2007) that once a book is sold, the current owner is free to do what he wants with the used copy. Authors and their agent publishers cannot prevent the sales of competing used works, nor do they receive any payments for those sales nor have any control over price.[5]

When we limit the analysis to new editions (by requiring the variable "new" to equal 1), as shown in the third column, the coefficient becomes much smaller (6.6%) and is no longer statistically significant. This reverses her conclusion that copyright increases the price of books, which I discuss in the main paper.

Reimers' Measure of Sales and her Welfare Results

Amazon does not provide data on unit sales, but it does provide a ranking of editions based on sales. Reimers scrapes Amazon data on editions to generate title sales estimates by month. Reimers takes an improvement in rank (moving to a lower number) as evidence of a sale, without explaining exactly how she does this.[6] By counting the number of improvements in rank, she

[4] I might note that if "collectables" are removed, defined here as editions with prices above $65, removing just 11 observations of a single edition (with prices of $98), excluding the pages variable, the price coefficient falls to 0.17 and is no longer significant.

[5] Obviously, used markets can influence markets for new products and might be of interest for some other questions about the market (e.g., see S. J. Liebowitz, "Durability, market structure, and new-used goods models" 72 (4), 816-824 *American Economic Review*). But it does not belong in an investigation of copyright's impact on the price of new books, which is the only market to which copyright applies.

[6] She does not explain whether any improvement in rank is counted, or only improvements of a certain magnitude. She doesn't tell us if she checks for rank changes once a day, once an hour, or once a week. She doesn't tell us if she tries to convert larger or smaller rank changes into larger or small changes in sales..

Electronic copy available at: https://ssrn.com/abstract=4547458

generates an estimate of sales for each edition which are then aggregated into title sales. Her (Amazon) sales data run from Sept 2011 to August 2012.

Using a change in rank as evidence of a sale has obvious problems, particularly when more than one sale occurs within a brief period of time, such as an hour, or when the impact of sales from prior days or weeks recedes and tends to reduce the rank, in contrast to current sales which improve the rank. These issues are likely to lead to undercounts for fast selling editions where multiple sales might show up as a single increase in the ranking. There is also the possibility that a rank my rise just due to normal fluctuations in the sales of other books sold on Amazon, providing possible overcounts of sales.

When we compare Reimers' measured sales with BookScan measurements (using 2012 as the year, but also examining 2011 as a check on the 2012 numbers) there are major differences, some expected and others not. Many of her listed Amazon sales numbers for individual titles are far lower than BookScan numbers, but some of that difference is simply because Amazon sales are less than that in the entire market. Reimers' attempts to adjust her Amazon sales values to cover the entire sales market by comparing BookScan values to her measured Amazon values for individual titles and finding the average BookScan advantage in sales relative to Amazon, which she measures as 2.5. She then multiplies her Amazon values by this number in an attempt to provide sales values for the entire market.

Unfortunately, when calculating her 2.5 value she takes an unweighted average of all titles whereas an average weighted by sales is required to find the *overall* sales difference between BookScan and Amazon for this set of vintage titles. My analysis using a weighted average indicates that overall BookScan sales of vintage titles are much larger than her Amazon measurements (almost 6 times as large) compared to her 2.5 value.

But there are other crucial differences in sales measurements as well. Some of the difference between our measurements is that she fails to include some large selling editions in her analysis.[7] But it is more than just that. The major problem appears to be with her measurement of sales and in particular her undermeasurement of top selling titles, a disproportionate number of which turn out to be copyrighted titles.

The four top selling titles, representing 79% of all vintage title sales, have sales that are 16.6, 12.9, 7.4 and 8.8 times as large as what Reimers measured as Amazon sales, compared to her calculated adjustment of 2.5. The top three titles are CR and the fourth is PD. Because Reimers uses her sales values when calculating the consumer and producer surplus of titles, she is going to greatly underestimate the CS and PS for those top four titles.

[7] For example, the largest selling edition of Gone with the Wind, ISBN 9781451635621, responsible for about half its sales, is missing in Reimers' data. Similarly missing are the largest selling editions of Babbit (9781593082673), Lost Horizon (9780062113726), and Main Street (9781593083861). Also missing are the second-best selling editions of All Quiet on the Western Front (9780449911495) and The Good Earth (9780743272933), the highest selling title in 2004.

Electronic copy available at: https://ssrn.com/abstract=4547458

For example, Reimers estimates the Amazon sales of the top selling title at 2,953 units. Multiplying that sales value by 2.5 brings the presumed market sales of that title to about 7,500 units. But BookScan lists the total sales for that title as 48,994 units sold units. The main reason for this difference is that Reimers' method indicates that the best-selling edition (9780449213940) of that title sells 2331 units (which should be multiplied by 2.5) whereas BookScan shows that edition selling over 42,000 units in 2012 (and over 56,000 in 2011). Reimers' adjusted values are thus off by a factor of almost 7 for this title. This means that the CS and profits generated by this title are much larger than Reimers' calculations.

It is also the case that there are 57 titles that Reimers reports as having positive sales although BookScan, which includes Amazon sales, reports no sales for those titles. In one instance Reimers reports (adjusted) sales of 175 units whereas BookScan reports sales of 1 unit. This would seem to support a view that Reimers' method which undercounts sales of top selling titles also overcounts sales for slow or zero selling titles.

In the main text's Table 10, I adjust for errors of this type, large and small. I calculate the ratio of BookScan sales to Amazon sales after Reimers' "2.5" adjustment. Most large selling titles have ratios of greater than 1 and most slow selling titles have ratios of less than 1. I multiply the monthly values of sales in Reimers' data set for each edition by these ratios and then allow the rest of Reimers Stata "do" file to perform its calculations. This simple adjustment provides results that overturn her conclusions about the impact of copyright on social welfare.

These are not the only adjustments that should be made to her data, however.

The role of Project Gutenberg in her results

The main reason that Reimers concludes that copyright harms welfare, even ignoring her erroneous sales values, is that her sales of PD titles include not just copies of titles (including digital ones) from regular publishers, but also digital PD titles that are made freely available from organizations such as Project Gutenberg and the Google Books Project. Her data are mainly from 2012, after the growth of eBooks and eReaders.

The first row of Table 52 (and also Table 10 in the main paper) provides the main results found and reported by Reimers, where the total surplus is 50% higher for PD titles than for CR titles.[8] The second-row reports results from her method and data if downloads from Project Gutenberg are removed from the analysis. The surplus generated by physical PD titles ("no Gutenberg") is seen to be very much smaller than that from CR titles. Thus, without Project Gutenberg her results would be completely reversed, and that is ignoring the adjustments I made when fixing her sales values for both CR and PD titles, in the previous section of this Appendix. Thus, it is

[8] My number $12,532 for the PD consumer surplus in row 1 in Table 9, row 1 is correct even though she provides a slightly different value of $12,278 on her page 275. We know this because on page 280 she provides a difference in CS between PD and CR titles that matches my value of 9982, and we have the same value for CR CS ($2550).

Electronic copy available at: https://ssrn.com/abstract=4547458

essential that her Gutenberg numbers be correct if her conclusions from her paper are to be considered internally consistent.

Reimers scraped numbers from Project Gutenberg to determine the number of downloads for titles. I have no issue with these numbers. The question is what those numbers mean.

Table 52: Impact of Project Gutenberg on Reimers' Results

		Avg CS per title	Profit	Total Surplus
(1) Reimers Results	CR	$2,550	$5,460	$8,010
	PD	$12,532	$0	$12,532
(2) No Gutenberg	CR	$7,104	$4,665	$11,769
	PD	$2,347	$0	$2,347
(3) Assume Half Gutenberg	CR	2593	5646	$8,239
	PD	6824	0	$6,824
(4) Adjusted sales, Half Gutenberg	CR	$6,409	$22,377	$28,786
	PD	$7,091	$0	$7,091

There are two problems with Reimers' Gutenberg data. The first has to do with the audience. First, anyone in the world can download the Gutenberg titles.[9] Since Gutenberg was the first and best-known site of its kind (begun in 1971) it is likely that readers in other English-speaking countries were likely to download titles from this site. The English-speaking populations of the UK, Nigeria, Canada, and Australia sum to about half of the US population. A much larger share of the world speaks English as a second language, dwarfing the US population, and many of these individuals read English well enough to read English language books.

Additionally, about 20% of the available title on Gutenberg are in languages other than English (only about 1% are in Spanish, the second leading language in the US), indicating that the population of downloaders of the Gutenberg site goes well beyond American English speakers. Many Gutenberg downloaders likely live in poorer countries, making free Gutenberg downloads particularly appealing, although the downloader/reader would need to have Internet access for the download and a device with which to read the books.

The point of the above paragraphs is to make clear that the number of Gutenberg downloads might strongly overstate the number of these titles downloaded by Americans. The welfare values generated in the above table (and in Reimers' paper) are for the American market only. It is obvious that we do not want to include non-Americans when measuring the value of Project Gutenberg when non-Americans are not included in our other measurements of surplus and profits. Therefore, Reimers' calculations, by treating all Gutenberg downloads as belonging to

[9] Project Gutenberg states: "Non-US persons are advised to check copyright laws of their country before accessing any eBooks or other content from Project Gutenberg." https://gutenberg.org/policy/terms_of_use.html.

Electronic copy available at: https://ssrn.com/abstract=4547458

Americans, overstate the value of Project Gutenberg downloads in the American market to the extent that the readership is not Americans.

Second, the values created by Gutenberg downloads should be lower than from purchases since it includes people with any positive value for the titles. Purchasers of books put forward a greater expenditure of time and money to purchase a book than downloaders from Project Gutenberg. Treating each download as the equivalent of a purchase seems likely to greatly overstate the value of PG downloads. It is even possible to download multiple Gutenberg texts at one time (or the entire library at once), reducing the average value of downloaded titles even further.

Given that non-Americans are likely to represent 50% or more of Project Gutenberg's downloaders, and given that the value of downloads should be treated as less than the value of purchases, it seems conservative to assume that the number of Project Gutenberg downloads should be halved relative to the measurements put forward by Reimers.

If we merely adjust the Project Gutenberg values in this manner, we arrive at the welfare results shown in row 3 of Table 52. This adjustment to Project Gutenberg values, by itself, is sufficient to alter Reimers' results to the opposite conclusion she reaches, although the advantage of CR titles is not large. In other words, it indicates that CR titles provide greater average value than PD titles. If you then make the further adjustment of fixing Reimers' sales values discussed in the previous section of this Appendix, you get the results in row 4, where CR titles provide overwhelmingly greater value (4x) than do PD titles.

Electronic copy available at: https://ssrn.com/abstract=4547458