Digitization and the Demand for Physical Works: Evidence from the Google Books Project

Abhishek Nagaraj
UC Berkeley-Haas
nagaraj@berkeley.edu

Imke Reimers
Northeastern University
i.reimers@northeastern.edu

April 12, 2021

Abstract

Digitization has allowed customers to access content through online channels at low cost or for free. While free digital distribution has spurred concerns about cannibalizing demand for physical alternatives, digital distribution that incorporates search technologies could also allow the discovery of new content and boost, rather than displace, physical sales. To test this idea, we study the impact of the Google Books digitization project, which digitized large collections of written works and made the full texts of these works widely searchable. Exploiting a unique natural experiment from Harvard Libraries, which worked with Google Books to digitize its catalog over a period of five years, we find that digitization can boost sales of physical book editions by 5-8 percent. Digital distribution seems to stimulate demand through discovery: the increase in sales is stronger for less popular books and spills over to a digitized author's non-digitized works. On the supply side, digitization allows small and independent publishers to discover new content and introduce new physical editions for existing books, further increasing sales. Combined, our results point to the potential of free digital distribution to stimulate discovery and strengthen the demand for and supply of physical products.

> We're absolutely certain that Google Book Search is making a difference to sales of the backlist... It's the publishing equivalent of being able to walk around a car, look under the bonnet and kick the tyres before making the decision to purchase.
>
> Cambridge University Press (Google, 2007)

1 Introduction

Digitization and the advent of the Internet have dramatically affected offline markets for information goods such as books, movies and music (Brynjolfsson et al., 2003; Forman et al., 2009; Greenstein et al., 2013; Waldfogel, 2017). The internet shapes physical markets by providing an alternative channel through which content can be consumed, often at very low cost or for free. This phenomenon raises the question of whether and how free digital distribution affects the sales for physical versions of information goods. A large cross-disciplinary literature has extensively studied this question, largely focusing on the cannibalizing effect of piracy on legal demand in markets for music and movies. The literature is extensive, and existing studies have mostly found that low-cost or free digital provision reduces or does not affect physical sales. For example, in music markets, "almost all empirical studies ... find that file sharing has caused a substantial decrease in music sales" (Smith and Zentner (2016), p.6) and the same holds true for movies, albeit to a lesser extent.

While the finding that free digital distribution tends to cannibalize (or not increase) physical sales seems relatively well established, less attention has been paid to the possibility that free digital distribution can also enhance search and discovery, thereby stimulating offline demand. Such an effect could be particularly salient in the market for books, where digital distribution can be accompanied by technologies that allow consumers to search through the full-text and discover (and buy) new content that would be otherwise hard to find (Ellison and Ellison, 2018). On the supply side, search-enabled digital distribution might allow distributors of physical books to discover and distribute new content, further increasing physical demand. Despite this possibility, past work has largely focused on markets where search-enabled discovery has traditionally played a less important role (music and movies), and has therefore not found any positive effects of digitization on physical sales. Investigating the effects of such search-enabled digital provision is important,

given the large size of the publishing industry, and because discovery through digital distribution might become increasingly relevant in other settings as well. Motivated by this omission, we examine whether it is possible for free digital distribution of books to increase rather than depress physical sales, especially when accompanied by a full-text search technology.

We begin our analysis by developing a simple theoretical framework that incorporates the discovery mechanism when considering the role of free digital distribution in shaping demand for physical books. The framework clarifies that while the net effect of book digitization is ambiguous, sales could increase if the discovery channel can compensate for the cannibalization of physical sales via the digital channel. While the cannibalization effect likely depends on the quality of the digital product, the framework especially clarifies that the discovery effect should be stronger for less popular books, should apply to non-digitized books by a digitized author, and should be muted for those who already had access to alternate search technologies prior to digitization. Finally, the framework considers the effect of digital access on the supply of new editions and suggests that digitization could increase the availability of follow-on editions, especially from smaller, independent publishers.

The heart of our study empirically analyzes the effects of a prominent, search-enabled, free digital distribution program: the Google Books digitization project. Launched in 2005, Google Books is one of the landmark projects of the digital age, with commentators likening it to a "modern-day Library of Alexandria" (Somers, 2017). Google Books did not just scan a book's textual material but also made it searchable via optical character recognition (OCR) technology through its "Google Book Search" feature (referenced by Cambridge University Press in the epigraph). Further, a large portion of the Google Books corpus included less well-known and older books (including public domain content) that are of significant consumer interest but have become forgotten over time. Google Books' ability to search through the voluminous set of printed works and locate those that pertain to a specific topic is likened to helping consumers and distributors locate a needle in a haystack. These features make Google Books a prime candidate for our study given that we are interested in examining the potentially positive effect of digital distribution on physical sales by helping consumers discover new works. Further, our setting is also of policy interest because Google Books' role in cannibalizing sales of existing editions was under significant litigation and has even been presented in front of the US Supreme Court.[1]

Investigating the empirical effect of free digital distribution on retail sales has traditionally been challenging due to data and identification challenges. An ideal experiment would randomly provide free, search-

[1] The Supreme Court ultimately declined to hear the case.

able, digital copies of a subset of books and link this variation to changes in physical sales before and after digitization. We tackle the empirical challenges through a unique natural experiment leveraging a research partnership with Harvard's Widener Library, which provided books to seed the Google Books program. The digitization effort at Harvard only included out of copyright works, which – unlike in-copyright works – were made available to consumers in their entirety. This allows us to fairly assess the tradeoff between cannibalization (by a close substitute) and discovery (through search technology). Owing to the size of the collection, book digitization (and subsequent distribution) at Widener took over five years, providing significant variation in the timing of book digitization. Further, our interviews with key informants suggest that the order of book digitization proceeded on a "shelf-by-shelf" basis, driven largely by convenience. While their testimony is useful to suggest no overt sources of bias, our setting is still not a randomized experiment, so that we perform a number of checks to establish the validity of the research design and address any potential concerns.

We obtained access to data on the timing of digitization activity as well as information on a comparable set of never-digitized books, which allows us to evaluate the impact of digital distribution on demand for physical works. Specifically, we combine data from three main sources. First, we collect data on the shelf-level location of books within the Harvard system between 2003 and 2011 along with information on their loan activity. Since most books are never loaned, our analyses focus on 88,006 books (out of over 500,000) that had at least one loan in the sample period (and are robust to using a smaller sample of books with at least one loan before the start of digitization). Second, for a subset of 9,204 books (in English with at least four total loans), we obtain weekly US sales data on all related physical editions from the NPD (formerly Nielsen) BookScan database. The sales data must be manually collected and matched, which restricts the size of this sample. Finally, we are interested in the effect of digital distribution on physical supply through the release of new editions. Accordingly, we also collect data from the Bowker Books-In-Print database on book editions and prices, differentiating between established publishers and independents. We use these combined data and the natural experiment we outlined to examine the effects of free digital distribution on the demand and supply of physical editions. Our panel data structure allows for a difference-in-differences design that can incorporate time and, notably, book fixed effects, increasing confidence in the research design.

The baseline results suggest that rather than decrease sales, the impact of Google Books digitization on sales of physical copies is positive. In our preferred specification, digitization increases sales by 4.8 percent and increases the likelihood of at least one sale by 7.7 percentage points. We confirm our findings in a series

of robustness checks and tests of the validity of the research design. First, in addition to book and year $\times$ shelf-location fixed effects, we also incorporate time-varying controls at the book level such as search volume from Google Trends and availability on alternative platforms like Project Gutenberg. Second, we provide a number of subsample analyses dropping certain books that raise concerns about the exogeneity of their timing, including limiting the data to only public domain and scanned books. Third, we create a "twins" sample that consists of pairs of scanned and unscanned books adjacent to each other in the library shelves and hence covering the same subject. Finally, we also collected data on Amazon reviews for a set of books in our sample as an alternate measure of physical demand. All results are largely in line with the baseline results.

We then evaluate and find that the increase in sales is likely driven by the discovery channel. Consistent with our theoretical expectations, digitization largely increases sales for less popular books, and these positive effects disappear for more popular books. Further, digital distribution increases sales for non-digitized works of an author with at least one digitized title in our sample. The significant and positive effect on sales for this sample suggests spillovers on demand across different works by the same author that are likely driven by discovery effects.

Next, we examine the importance of the discovery and substitution channels by studying two parallel settings where one of these two effects is muted. First, we investigate the effects of digitization on loans within the Harvard system. In this setting, the discovery effect is muted since Harvard students and professors already had access to alternate discovery mechanisms through library services. In this setting, digitization reduces rather than increases demand, measured as loans within Harvard. Second, we investigate an alternate sample of 11,166 recently published in-copyright books sold on Amazon with and without the "Search Inside the Book" (SITB) feature enabled. This feature enables customers to browse a book in "snippet" view and search the full text before purchasing, but it does not provide the entire text for free, muting the substitution effect. We find that SITB status is correlated with higher demand, suggesting that enabling search is important to increase offline demand via digital provision.

Finally, we test the prediction that digital provision increases the supply of physical works by enabling publishers to introduce new editions, further boosting sales. Regressing the flow of new physical editions on digital availability, we find that digitization increases the number of new editions for books. Even though these estimates come from a sample of public domain works (where publishers do not need to license content to introduce a new edition), these results suggest that digitization can help boost the supply of physical editions, at least under certain conditions. Supporting the discovery mechanism, this effect is

largely driven by independent publishers, who presumably have fewer resources for finding good texts than larger publishing houses and university presses. When we account for this increased supply of new editions, we find that this channel is responsible for about 50% of the overall physical sales effect, with the remaining half coming from increased demand for existing editions. These results suggest that digital distribution can stimulate sales of published content, both through increased supply of new editions and through increased demand for existing editions. Free digital distribution can also facilitate the entry of smaller publishers who can use digitized content to discover and distribute new content, shaping competition in the market for physical information goods.

Our study makes three key contributions. First, our results stand in contrast to a large literature has examined the effects of free digital distribution on retail sales in the content industries and has largely found that digital distribution harms or has no effect of sales of physical copies. Our contribution is to show that, when accompanied by search technologies that enable the discovery of new products, digital provision might have the capacity to boost rather than cannibalize physical sales. A strand of this literature has studied the effect of cross-channel substitution in the market for books, but does not specifically focus on the effects of digital distribution on retail sales. Our work is closest to Chen et al. (2019) who study the effects of ebook availability on physical sales and do not find any negative or positive effects. Second, academic literature has investigated a variety of mechanisms through which consumers can discover new content. We show that such discovery effects can be strong enough to compensate for the cannibalization effects of digital distribution. For example, our study is related to Zhang (2018), who shows positive effects of lifting sharing restrictions on sales, but does not study the interactions between (free) digital distribution and physical consumption, nor does it study the market for books. Finally, we add to past work that has studied the impact of legal restrictions such as intellectual property on the supply of works by encouraging reuse and diffusion of content. We show that digital distribution can help distributors discover new content and boost physical distribution, highlighting another potential channel through which digitization can encourage physical supply. Since these effects are largely driven by independent publishers, our results are among the first in the literature to point to the effects of digital distribution in shaping offline competition and entry.

The paper proceeds as follows. We begin by providing a brief overview of the academic literature and describing the background of the Google Books project in Section 2. We then describe our data and research design in Section 3, followed by a description of the main results, robustness checks and an exploration of the discovery mechanism in Section 4. We conclude with policy implications in Section 5.

2 Related Literature and Theoretical Argument

2.1 The Impact of Digital Distribution on Physical Demand

Our work is closely related to the literature that has looked at the impact of free or low-cost digital distribution of information goods on demand for physical alternatives. This work has largely studied the effects of piracy on the markets for movies and music (See Smith and Zentner (2016) for a review).

A number of papers have looked at the effect of illegal online distribution in the form of piracy on sales of legal music and movies. The majority of this work looking at the impact of free distribution via file sharing on music sales finds a negative effect (Bounie et al. (2006); Rob and Waldfogel (2007); Zentner (2005, 2006); Rob and Waldfogel (2006) see Danaher et al. (2014) and Oberholzer-Gee and Strumpf (2010) for a review), although some work on the movie industry (Bai and Waldfogel, 2012; Danaher et al., 2010) suggests a less pronounced effect. More recently, scholars have also looked at legal forms of cheap, digital distribution such as online streaming and found that it too tends to depress sales in other channels (Yu et al., 2018; Aguiar and Waldfogel, 2018). In these industries, the extant literature does not tend to find that digital distribution stimulates demand for physical products, perhaps because digital distribution does not explicitly improve the information environment.

Insights from the literature on music and movies might be less relevant to the market for books given one key difference: Digitization has the potential to allow for the scanning and searching of the entire text of the book. This feature permits a much deeper match between content and a customer's preferences. Given that match quality between content and readers is quite important in the market for books (Ellison and Ellison, 2018), a full-text search and digital distribution technology that matches readers to books can in fact increase demand, even if books are distributed online for free. Demand would increase by helping readers locate books that they would otherwise not know about or want to read. This effect can increase sales of physical copies, especially if readers find it cumbersome to read lower quality scans in an online venue (such as Google Books) or in cases where readers do not have access to the full text through the digital medium (such as through "snippet" access).

Scholars have studied aspects of the market for books such as price dispersion (Ellison and Ellison, 2018), comparing product variety in online and offline formats (Brynjolfsson et al., 2003), substitution between used and new books (Ghose et al., 2006) or platforms for books access (Baye et al., 2015), but the impact of digital distribution on physical sales and especially the role of full-text search are less well understood. The few studies that do look at the impact of digital distribution on physical sales in the publishing

industry study contexts where digital distribution is provided without the added benefit of full-text search. For example, Chen et al. (2019) consider the impact of e-book distribution on physical book sales, finding no effect. Similarly, Forman et al. (2009) find that digital and physical distribution channels for books are largely substitutes when considering Amazon.com. A key question that motivates our paper remains unanswered: What is the effect of digital distribution – combined with full-text search technology – on physical demand in the market for books?

Answering our research questions is important, not only because it helps make theoretical progress on the literature on cross-channel substitution, but also because of its industry and policy relevance. The net revenue of the book publishing industry in the US in 2019 was more than thrice as large as the music revenues ($25.9B as compared to $7.3B; AAP 2020 and RIAA 2020, respectively). Further, the advent of the internet, platforms like Amazon.com and mass digitization projects like Google Books have presented numerous questions about balancing physical sales, digital distribution (via e-books) and mass digital distribution (via projects like Google Books) for firms in the publishing industry. In fact, legal debates have analyzed the specific case we focus on (the Google Books project) and debated whether it increases or decreases sales (Samuelson, 2009). Our research provides quasi-experimental evidence to policy-makers and legal scholars, who have largely relied on anecdotal and theoretical data up to this point.

2.2 Theoretical Argument

How might digital distribution increase the sales of physical books? For end customers, the question of whether the digitization of books increases or reduces demand for physical works depends on two counteracting forces. The first is the substitution effect of digital distribution as a competitor for existing, physical products as studied in the literature. Some consumers who would otherwise consume physical copies will switch to digital versions, driving the substitution effect. This is likely to happen when a consumer's search costs are low, and when she has a taste for digital consumption, and it may be less relevant when books are cumbersome to read in the online format (e.g., on Google Books).[2] However, Google Books might stimulate a discovery effect due to increased awareness and searchability (see appendix Figure D.1.) Specifically, some consumers may start consuming the physical version for the first time, after digitization lowers search costs for books. This is likely to happen if they were made aware of a book through Google Books' search engine and prefer to purchase physical copies rather than read online. This second mass of consumers will drive the discovery effect. The net effect of digitization is ambiguous and depends on the magnitude of these

[2]Note that this tradeoff is different from, say, MP3s vis-a-vis CDs, where MP3s may have even delivered a better experience than their physical counterparts.

two margins.

The tradeoff between substitution and discovery further differs for different margins of books and consumers. Notably, for popular books, already well-known to consumers (e.g. *The Wealth of Nations*), the substitution effect is likely to dominate. On the other hand, obscure books are likely to benefit from discovery, and unlikely to face the costs of substitution. The effect of Google Books on demand should therefore be more positive for less popular books. In addition, if consumers discover a particular author through a digitized copy, they might also seek out other books by the same author, even if these have not been digitized. Therefore, digitization might lead to an increase in physical sales for non-digitized works of a digitized author as well.

Further, when the discovery channel is muted, the positive effects on demand should reduce or disappear altogether. For instance, for consumers within Harvard, who already benefit from access to search technology (through Harvard's librarians and internal catalog system) the substitution effect is likely to dominate the discovery effect. Therefore, when considering loans within Harvard, the effect of digital distribution is likely much smaller, and even negative. On the flip side, when a digital platform provides access only to the search function, but not the entire text of the book (as is common with "snippet view"), we expect the positive effect of demand to remain strong. Our empirical analysis sheds light on these predictions as well.

Digital distribution might also shape the supply side. Specifically, imagine that publishers publish any content that nets them positive revenues as long these are greater than the fixed costs of locating and licensing materials of interest to their audience. Digitization lowers search costs and helps publishers identify interesting content that typically would be unknown to them, making it more likely that they will produce a new physical edition for a book. These dynamics are especially likely to be at play when the underlying content is not in print (and publishers face no competition) or when it is in the public domain and free to license. Similar dynamics could apply to in-copyright content if there is an active market to license out-of-print or less popular content for existing license holders, albeit to a lesser extent. At the same time, digitization could also increase competition or lower prices, which might reduce publishers' profits per edition, lowering the likelihood that they will increase the supply of new editions. If the competitive or price effects are minimal, we expect that digital provision will increase the supply of new editions, although this remains an empirical question. Further, any positive effects on supply should be especially relevant for small and independent publishers, who face higher costs of locating content since they do not have an existing catalog or network to source such material.

To summarize, our theoretical predictions are threefold. First, digital distribution allows consumers to search for topics they are interested in and discover works previously unknown to them. Second, if the digital medium offers a poor substitute for a physical book, demand for physical works is more likely to increase. Finally, digital provision will also allow publishers (especially small and independent ones) to identify new material and introduce new editions.

Our theoretical arguments speak to past work on the impact of digital distribution on demand and supply for physical products. On the demand side, our arguments about the role of digital distribution in enhancing consumer discovery add to work on the effects of digital technology on the diversity of consumption patterns (Brynjolfsson et al., 2006; Kumar et al., 2014; Holtz et al., 2020). This work shows that digital distribution channels tend to change consumption patterns by helping consumers discover more novel and niche products. Related work has looked at the complementary effects of news content sampled via social networks, which drives traffic to news websites by helping consumers discover specific news articles (Chiou and Tucker, 2017; Sismeiro and Mahmood, 2018). However, this literature focuses on how digital channels change consumption patterns and has not explored how online access coupled with access to digital search and discovery tools affects demand for and supply of the same product in physical form. Our theoretical and empirical analyses extend this work in this direction.

On the supply side, past work has looked at other channels through which access to existing work improves the supply of new content. Most notably, the literature in copyright and digitization shows that free access to past work can often stimulate follow-on production of knowledge (Watson, 2017; Reimers, 2019; Heald, 2007). For example, the (copyright-related) digitization of magazine content improved the quality of content on Wikipedia (Nagaraj, 2018). Further, existing literature also suggests that digital access can be particularly helpful for smaller players and stimulate entry (Nagaraj, 2020; Zhang, 2018). However, whether or not digital distribution itself can improve the supply of physical products and whether it benefits smaller or larger players on this margin remains unexamined.

3 Setting, Data and Research Design

3.1 The Google Books Project: A Brief Background

The Google Books project (originally known as the Google Print Library Project)[3] was announced by Google in December 2004. At the project's inception, Google partnered with Harvard University's library (along with a few other key partners) to digitally scan books from their collections. Soon – usually just

[3] https://googleblog.blogspot.com/2004/12/all-booked-up.html

a few weeks – after these works were scanned, they were made available on the Google Books website for the general public. The site provided access to the full-text of public domain books (including books published in the US before 1923) but only a "snippet" (i.e. limited) view for in-copyright material. Further, an important feature of the site was the ability to search through the entire text of all scanned books.

Soon after its launch, the Google Books project was met with staunch opposition from the Authors Guild and the Association of American Publishers, who filed class action suits against Google for copyright violation.[4] Hearkening to arguments in the literature on cross-channel substitution in music and movies, large groups of authors and publishers expressed concern about the possibility that digital distribution could cannibalize physical sales. In an online statement, the Authors Guild claimed that "Google Books can create a very real negative economic impact on the books it has digitized ... rather than drive researchers to buy books, readers for many books can find all they need on Google Books."[5] Google Books' major defense was centered on the idea that browsing books may promote the downstream sales of digitized material.[6] The argument here was that Google Books' digitization efforts "increase[d] the visibility of in and out of print books, and generate[d] book sales,"[7] and that it was "designed to help you discover books, not read them from start to finish."[8] Some publishers subscribed to Google's argument and were not opposed to the project – in fact, some, like the Cambridge University Press, adopted it for their back catalog, although the overall opposition to the Google Books project remained. The suits were eventually settled (publishers) or rejected (authors). As a result, the upshot is that "somewhere at Google there is a database containing 25-million books" that is inaccessible to the general public (Somers, 2017). In fact the real number is probably higher: A blog post by Google in 2019 reports that Google Books has digitized over 40 million books.[9]

While Google Books was not the only project digitizing works, it was both the most comprehensive and the most publicized. Two of the largest related projects digitizing public domain works are Project Gutenberg and the Hathi Trust. Project Gutenberg, which made the first digitized public domain work available in 1971, offers just under 65,000 works as of March 2021. The Hathi Trust, which was founded in 2008, offers almost 17.5 million total volumes (8.4 million book titles) by the same date. Moreover, Google Trends reports that the Internet search volume for Google Books was over twice as large as that for Project

[4] See Samuelson (2009), and https://googleblog.blogspot.com/2008/10/new-chapter-for-google-book-search.html.

[5] https://www.authorsguild.org/where-we-stand/authors-guild-v-google/, Accessed April 4, 2019

[6] See Authors Guild vs. Google (SDNY 2013), https://h2o.law.harvard.edu/collages/34596 for more information on the case.

[7] See http://googlepress.blogspot.com/2004/12/google-checks-out-library-books.html.

[8] https://web.archive.org/web/20041214092414/http://print.google.com/

[9] See https://www.blog.google/products/search/15-years-google-books/.

Gutenberg between 2004 and 2011 (our period of study), while no search volume was reported at all for the Hathi Trust, suggesting that it was not very popular by 2011.

3.2 Google Books and Harvard Libraries' Natural Experiment

Given the unclear legal environment around digitization and copyright when the project began, and due to concerns about potential copyright challenges and bad publicity, Harvard's participation in the Google Books project was limited to out-of-copyright works from Harvard's prestigious Widener Library.[10] Under the Copyright Term Extension Act of 1998, it is clear that works published in the United States before the year 1923 are in the public domain. Therefore, Harvard provided US books published before this year for scanning. Since this cutoff date would not change until long after the digitization was completed, books from after 1923 were not digitized. Different cutoff dates were applied to international books in determining their inclusion in the scanning effort.

The digitization effort proceeded as follows. Google set up a scanning facility in the Greater Boston area to process the books from the Harvard libraries. For the purposes of the scanning effort, Google Books was assigned a special library patron code, and books were "loaned" to Google under this special code to be taken to the scanning facility. Once the book was scanned, it was returned to the library and made available on the Google Books website after a short delay, usually within a few weeks (personal communication, December 2011). We impute a book's scan date based on the checkout date at Harvard. Since a book could have been digitized at a library other than Harvard, it is possible that this date does not accurately reflect when a book was first made available on Google Books.[11] However, since Harvard was one of the first libraries to seed books for the Google Books project, our scan dates are likely to be representative of the first time a book was available in digital form on Google Books.

Google Books took over five years (from 2005 to 2009) to complete its large-scale scanning project at Harvard. In our baseline analysis, we rely on the variation in the timing of the scanning project across books to estimate the impact of digitization on eventual readership and sales, along with book and year fixed effects. Further, personal communications with key players in the digitization effort and an analysis of book locations within Widener (that we present later in this paper) indicate that the order in which books were scanned was primarily driven by convenience rather than an explicit selection mechanism. We conducted a number of interviews with university officials involved with the Google Books project, including a key

[10] Other libraries involved in the early phase of digitization, such as at Stanford and the University of Michigan, did not act on these concerns and also offered in-copyright works for digitization. Those works were then made available as "snippets" – showing only small excerpts of the text – on Google Books.

[11] If this were the case, we would be less likely to find any effects in our empirical analysis.

official at Harvard University who was responsible for administering the collaboration with Google. In our interview, he clarified that books went to the Google scanning facility "shelf by shelf" and that it was a "very fast, continuous flow" and "bulk work" and that Harvard did not "look at it in terms of subject or anything else" (interview with authors, Jan 15, 2021). He reiterated that since a large number of books was involved, there was "absolutely no selection on any base" other than the script of the books (with books in roman script prioritized). This interview is reassuring, because it suggests that there was no intentional effort on the part of Google or Harvard to prioritize books for scanning. We heard similar anecdotes from other university officials, for example at the University of Michigan, who confirmed that there was no attempt to select books for early digitization on their part. While these qualitative reports are reassuring, it is still possible that there are unintentional sources of selection that could affect our estimates. We will investigate these concerns through our quantitative data analysis.

3.3 Data

The data we obtain from Harvard contain the entire record of over 250,000 books from the Harvard libraries' holdings that were scanned, as well as a similar number of works published between 1923 and 1943 that were not scanned. We use this underlying set of books as the basis for constructing our dataset from three separate sources. First, we possess proprietary checkout data, which allows us to infer the date when the book was checked out by Google Books for digitization, as well as total loans within Harvard. Using these data, we construct our baseline sample, which consists of all 88,006 books that were checked out at least once between 2003-2011.[12] Our sample of 88,006 books includes 37,743 (43%) that were scanned between 2005 and 2009 and 50,263 (57%) that were under copyright and not scanned. Its composition of subject areas is representative of works available at Google Books: about 9% of books in US History, 5% in Economics, and 3.5% each in British Law, Philosophy, Slavic Studies and American Literature.

Second, we obtain access to NPD (formerly Nielsen) BookScan, which provides sales information for printed books. NPD tracks book sales using scanner data from a large panel of retail booksellers including major bookstore chains, discount retailers such as Costco, and major online retailers like Amazon. They claim to track about 85% of total retail sales, although these data do not capture e-book sales.[13] Because our data from Harvard do not contain global unique identifiers (i.e., ISBNs), we (and a team of research assistants) manually search NPD BookScan for each book title to find suitable matches, aggregating sales

[12] Since this sample definition might be problematic because it conditions on one of our dependent variables, we examine robustness of our analysis to using only books with at least one loan *before* digitization and show that all results carry through.

[13] See Berger et al. (2010) and https://tinyurl.com/y94qpsqt, accessed June 26, 2018. The lack of e-book data does not limit this study significantly. Waldfogel and Reimers (2015) report that during the time of our study, e-book sales never make up more than 13% of the market.

of all hardcover and paperback editions for each title by calendar year. Given the tedious data collection process, we search for sales data for the subset of all English-language books in the underlying dataset with at least four loans, for a total of 9,204 titles, or 10.5% of the original titles.[14] Of these, 3,267 books (36%) are scanned for Google's digitization project.

Third, we collect data on the number of in-print editions of all works from the Bowker Books-in-Print database. This database tracks all registered editions of a particular work that are available in print. We match the 88,006 books in our sample to this database, finding matches for 25,719 unique titles with in-print editions.[15] Combined, the Harvard libraries data on book digitization and loans, the NPD BookScan data on book sales, and the Bowker Books-in-Print database on editions allow us to characterize the impact of digitization on the demand for physical works within Harvard (loans) and in the market (sales), as well as on their availability. This is, to our knowledge, the first dataset that matches the digitization status of works with data on their sales and in-print status.

We organize the data into a balanced panel at the book-year level between 2003 and 2011. Of the 37,743 scanned books that we analyze, 5,764 were scanned in 2005, 7,449 in 2006, 8,769 in 2007, 13,207 in 2008, and 2,546 in 2009. The variables of interest are summarized in Table 1 Panels A (book-level) and B (book-year level). In any given year, an average book sells about 554 copies, has 0.25 loans and adds 0.36 editions, although the median value for all three outcomes is zero. Over the entire sample, books have average sales of almost 5000 and are loaned on average 2.23 times.

The skewed nature of demand for the books in our sample leads us to study the impacts of digitization not only on the intensive margin – how many copies are consumed – but also on the extensive margin: will a work be read at all? Each year, books that are never scanned have an average annual probability of being sold of 16%, whereas those that are scanned have a probability of only 8.5% before their digitization and 24.1% after it. Similarly, books that are never digitized have a probability of 17.8%, while books that are digitized have a probability of 19.3% before their digitization but only 11% after their digitization. These differences are indicative of large potential impacts of digitization on demand.

[14] Because NPD BookScan does not explicitly list books with no recorded sales, we impute zero sales for titles that do not explicitly appear in the BookScan database. The results are robust to excluding these titles from the analysis.

[15] One reason we do not find more matches with the Bowker Books-in-Print database is because some works are not intended for a commercial audience, e.g. dissertations.

3.4 Testing the Validity of the Natural Experiment

Our difference-in-differences approach relies on the assumption that books digitized early experienced similar demand trends as books digitized later. In our analyses, we include book fixed effects and book-category $\times$ year fixed effects in different specifications in addition to testing for pre-trends in analyses of the annual impact of digitization. Still, since the timing of book digitization was not explicitly random, it is important to examine what possible sources of selection might exist. In this section, we identify challenges to the research design and motivate additional robustness checks.

First, we obtain the library call numbers for the titles in our sample, which helps us map a particular book to its exact location in one of twenty possible stacks within Harvard's Widener library. We are able to match about 81% of all scanned books to an exact stack within the library. Using these unique data, we examine the assertion that the timing of a book's digitization is largely based on its physical location. Figure 2 plots a heatmap of book digitization by library stack location (y-axis) over time (x-axis). The colors are based on the percent of books digitized in a given month as a fraction of the total number of books digitized between 2005-2009 in a given stack. The bluer the zone, the higher the percent of books from that stack that were digitized in that time period.[16] For example, 73% of scanned books in the B West stack were scanned in mid 2007, which is indicated by the bright blue spot on the heatmap for this stack. As the series of blue spots along the diagonal indicates, almost every stack has a single time period when a large percent of their books are digitized and this time period varies from stack to stack. Note that Harvard had three simultaneous streams of books going to the digitization facility, which might explain the lack of an explicit focal month for some of the stacks. The patterns in Figure 2 support our interpretation that books were digitized based on their stack location. One notable exception is Pusey 3, which has no single time period when a majority of its books were digitized. This is likely because Pusey 3 is a large stack, remotely located several floors underground the main floors, and its books are more likely to be stored in other remote locations. While conversations with insiders alleviate most concerns about selection, we also examine the robustness of our analysis to excluding books in this stack.

Next, we examine pre-digitization demand (i.e. in 2003-2004) for books based on the year in which they were scanned. If the timing of scanning is random, then book-level covariates should be unrelated to the timing of digitization. Accordingly, we regress the number of pre-digitization loans and sales on indicators for the year of digitization after accounting for subject and library location dummies (since we account for these variables in our regressions as well). The coefficients from these regressions that allow us to

[16] The stacks are sorted from bottom to top by the month in which the highest percent of books in their stacks were digitized.

compare the never scanned cohort (with has a coefficient of 0 by construction) with cohorts scanned in 2005, 2006, 2007 and 2008-09 are presented in Figure 3. Panels (i-iv) cover static measures of pre-digitization outcomes like the publication year, likelihood of being in the fiction category, pre-05 sales and pre-05 loans. Unsurprisingly, as Panel (i) shows, scanned books are published much earlier than never scanned books, but there is no difference between publication years across the different scanned cohorts. Panels (ii) and (iii) show no major differences between all cohorts in terms of subject matter or sales. However, looking at Panel (iv), books digitized in the first year of digitization (2005) seem to have a higher number of loans than books digitized later. This in itself is not problematic given that we employ book fixed effects. Regardless, we examine the robustness of our design to excluding all books digitized in 2005. Finally, given our book fixed effects design, we also evaluate *changes* in sales and loans prior to digitization as a measure of the "hotness" of a book that might be problematic for our research design. Panels (v-viii) examine the change in different measures of sales and loans between 2003 and 2004, and finds no significant differences across the different cohorts.

In sum, our analyses provide support for our qualitative interviews that suggested that the digitization process was driven by shelf location. However, there seems to have been some concerns for the Pusey 3 stack (that was digitized at different points in time) and for books digitized in 2005 (that have higher levels of pre-05 loans). As we will show, the additional tests motivated by these concerns are in line with our baseline results and further reinforce the research design.

4 Results

A first approach to examining how digitization affected sales of physical editions could take advantage of Harvard's decision to only digitize public domain books – those that were originally published before 1923 – and to leave copyrighted books (those from after 1923) untouched. We exploit the sharp cutoff around the publication year 1923 to examine whether sales of books published right before 1923 changed considerably compared to books published right after, once the digitization process has been completed.

Figure 1 illustrates how demand changed over the digitization period across different publication cohorts. It plots the share of the books in our sales sample that sold more copies in the two years after the digitization period (2010-11) than in the two years before the digitization period (2003-04), for each publication year for the 20 years before and after 1923. The figure shows stark differences in the likelihood of increased sales between digitized and non-digitized cohorts, with digitized books being much more likely to sell more copies after digitization. About 40 percent of digitized titles see a sales increase from 2003-

04 to 2010-11, compared to less than 20 percent of titles that were not digitized.[17] These cross-sectional differences suggest large possible effects of digitization. To quantify these effects, and to identify possible mechanisms, we take advantage of the staggered digitization across Harvard's entire catalog, as we describe in detail below.

4.1 Main Specification and Results

In our main specification, we compare the evolution of sales for titles that were scanned and made available on Google Books with that for titles that were not (yet) digitized in a difference-in-differences setting. Formally, we estimate equations of the form

$$Y_{it} = \alpha + \beta PostScanned_{it} + \gamma_i + \mu_t + \varepsilon_{it}, \tag{1}$$

where $PostScanned_{it}$ is an indicator that is 1 if book i has been made available on Google Books before year t, and γ_i and μ_t are book and year fixed effects, respectively. The dependent variable, Y_{it}, denotes book- and year-specific measures of demand. In a first set of baseline analyses, the dependent variable is the zero-inflated log-sales of all editions of the title ($\ln(sales_{it}+1)$). In a second set, we estimate a similar specification using a linear probability model (LPM) where the dependent variable is $1(sales_{it} > 0)$. That is, we examine the likelihood that a title will have any sales in a given year after digitization.

Table 2 displays the main results. All specifications show that market-wide sales increase after digitization. The first two columns report results from log-sales estimations, with book and year fixed effects (column 1), and book and year-library location fixed effects in column 2. Both specifications report statistically significant increases in the number of copies sold due to digitization, with an estimated sales increase of 4.8 percent ($=e^{0.0466}-1$) in the full model from column 2. At an average of 555 sales per book and year, the estimated increases are also economically significant. Columns 3 and 4 report similar effects from corresponding linear probability models. The full model from column 4 indicates a digitization-induced increase in the probability that a title is sold at all of 7.7 percentage points. Given the baseline probability of a sale, this suggests a nearly 50 percent increase in the probability of a sale.

We also allow for a flexible time structure by estimating the annual changes in a book's demand relative to its digitization year. Specifically, we estimate

$$Y_{it} = \alpha + \Sigma_z \, \beta_z (scanned)_i \times 1(z) + \gamma_i + \mu_t + \varepsilon_{it}, \tag{2}$$

[17] We show results from a more detailed regression discontinuity approach in appendix Section A.

where γ_i and μ_t represent book and year fixed effects, respectively, $(scanned)_i$ equals one for all books that were eventually scanned, and z represents the "lag," or the number of years since the book was first digitized, with 1 denoting the year of digitization.[18]

Panel A of Figure 4 illustrates the results from this specification, using both an OLS model with log-sales as the dependent variable (left figure), and the linear probability model where the dependent variable is an indicator that equals 1 if a copy of the book has been sold at all (right figure). Two points are clear from this analysis. First, there are no significant pre-trends in either specification, which provides support for the validity of our research design. Second, the positive effects on our sales measures seem quite persistent and long-lasting, and they kick in soon after digitization.

4.2 Robustness Checks

To bolster our baseline estimation, we present results from several robustness checks, including approaches to deal with potential endogeneity as alluded to above, selection issues, alternate sample restrictions, and alternate estimation specifications. We present results from these robustness checks in Table 3.

The first three columns examine the potential endogeneity of the timing of digitization. First, we include two additional control variables in our main regressions. While we only observe list prices for the physical editions and therefore cannot add price information beyond the title fixed effects, we can still control for potential changes in interest for a title over time by adding each title's annual Google search volume to the estimation. We obtain annual search volume for each title from Google Trends, which reports indices of search volume over time starting in 2004. The day with the highest search volume is normalized to 100, and we normalize search for the year 2003 (for which we do not have data) to 100 for all works.[19] To control for other changes in availability and attention, we further include an indicator that is one if the book has also been made available on Project Gutenberg – another major project attempting to digitize and make available all public-domain works.[20] We include these control variables in the estimation underlying column 1 of Table 3. The estimated effect of digitization through Google Books remains strongly significant and becomes even larger, now suggesting a digitization-related increase in sales of 6.5 percent.

Second, our analysis of the research design in Section 3.4 above suggests that books digitized in 2005 and those in the Pusey 3 collection might be systematically different from books digitized in later years or

[18] For books digitized before July in a given year, the lag variable equals one in the first year of digitization, while for books digitized in July or after, the lag variable is set to one in the calendar year after the year of digitization.

[19] The value of 100 for the 2003 normalization is irrelevant for our estimates given the use of year fixed effects.

[20] During our period of study, Project Gutenberg had about one third as many Google searches as Google Books, suggesting smaller likely effects.

located elsewhere. Accordingly, columns 2 and 3 of Table 3 drop these subsets of titles. Again, the baseline result remains highly significant, both without (column 2) and with (column 3) our new, time-varying control variables. In addition, we disambiguate the estimated effects by scan year in appendix Table D.2.

The next two columns of Table 3 address concerns about the sample in the main specifications. In column 4, we limit the data to books that were in the public domain and therefore digitized at some point during the study. This specification alleviates concerns that the unscanned books – those under copyright – may not be a good control group, and its results are consistent with the baseline results. In column 5, we limit the control group in a different manner. From our larger sample, we choose pairs of scanned and unscanned titles, located right next to each other on Widener's shelves as per their call numbers.[21] Two books that are located next to each other cover the same subject area and are usually quite similar. The scanned and unscanned books in this sample therefore cover almost identical subject codes and have very similar pre-digitization demand: On average, scanned books sold 394 copies per year and unscanned books sold 402 units per year (t-value of their difference = 0.07). Using this sample, we repeat both the baseline regression from equation (1), in column 5 of Table 3, and the flexible time structure regression from equation (2), in the bottom panel of Figure 4. The estimates from both specifications are very similar to those from the main sample, if not stronger.

The remaining columns of Table 3 explore the robustness of our results to the baseline model assumptions. The hyperbolic sine specification, $Asinh(sales_{it})$, in column 6 addresses potential issues from inflating the zeros in the dependent variable. In column 7, a Poisson estimation takes the countable nature of the sales variable more seriously; and in column 8, we estimate the likelihood that a copy of a book is sold at all in a binomial logit estimation. All specifications support our main quantitative and qualitative results. As the only exception, the Poisson specification provides larger but less precise estimates.

Finally, while the NPD Bookscan dataset covers the vast majority of all physical book sales, it does not tell us whether a sale is made on an online platform or at a physical bookstore, and the two channels may be affected differently. In a separate analysis, we obtain data on micro-level reviews at the Amazon platform from Ni et al. (2019). We use these reviews to proxy for a book's demand at Amazon, and we estimate the effect of digitization on these reviews. Our results suggest a digitization-related increase in annual Amazon reviews of 3.97 reviews, relative to an average of one review per year for books in our dataset.[22]

[21] For example, this approach drops all unscanned books that are located between two other unscanned books.

[22] See appendix Section B for more detail on the Amazon data and estimation strategy.

4.3 Digitization and Discovery

The discussion above suggests that digitization – when coupled with improved searchability – can in fact increase demand for physical copies, and our theoretical framework links these increases to facilitated discovery and variations in search costs. We highlight the potential discovery function of the full-text digitization through Google Books in three separate types of analyses. We examine whether the effect varies for books of varying popularity; we examine if the digitization of one book by an author also affects demand for the author's other books; and we separate the discovery and substitution effects by looking at settings where one of the two is muted.

Our first test estimates the effects of digitization on books of varying popularity. We divide our books into three groups according to their sales before Harvard's digitization effort began (i.e. in 2003-04): books with no sales (91.9% of the sales sample), books with one to 500 sales (3.1%), and books with more than 500 sales (5%). We then repeat our baseline estimations, interacting our post-scanned variable with indicators for each popularity group. The results are reported in Panel A of Table 4.[23] The first two columns repeat the first two columns in Table 2, and the remaining columns repeat the sample-based robustness checks from Table 3. We find a small but statistically significant positive effect (sales increase by 4.6%) in the first group (pre-sales=0), we observe large and significant positive effects (increases of over 30 percent) of digitization on books that had previously been relatively obscure (pre-sales between one and 500), and no statistically significant effect on the more popular works. Assuming that the discovery effect is less pronounced among books that are already well-known, these results provide first evidence that discovery plays a role in determining the effect of digitization.

In a second, somewhat stricter test inspired by Zhang (2018), we examine the effect of digitization on closely related books. If discovery plays a major role, then digitization could make a potential reader aware of an author, including their entire body of works, and therefore increase sales of all books by that author. We identify authors who have at least two books in our sample, and we estimate the effect of digitization of one book by an author on sales of the author's other books. Panel B of Table 4 report results from regressions that are again based on equation (1). The first two columns include only books that are protected by copyright, comparing never-scanned books by never-scanned authors with never-scanned books by authors with at least one scanned book. The other two columns include scanned books by these authors and separately estimate the effect on these books and on the author's other books. The estimates suggest large positive spillover

[23] We report results from robustness checks analogous to Table 3 in appendix Table D.3, and we explore more granular popularity cutoffs, which divide books with positive sales into quintiles according to their sales in 2003-04, in appendix Table D.4.

effects of the free digital provision through Google Books on other books, in size quite comparable to the direct effects.

Third, we attempt to examine the effects of the discovery and substitution effects more directly, in two settings that each mute one of the two effects. We first mute the discovery mechanism by estimating the effect of digitization on library loans through Harvard's library system. Because Harvard always had systems for discovery in place – including hundreds of librarians specializing in certain subject areas – patrons of the library likely experience less of a benefit from the searchability of the digitized versions of the books. Therefore, the substitution mechanism may outweigh the discovery effect in this setting. Table 5 reports results from variants of regressions of library loans on digitization. Showing results from the zero-inflated log-OLS model as well as a linear probability model, the first two columns imply that digitization decreases loans on average by about five percent, and decreases the probability that a book is checked out at Harvard at all in a year by 6.1 percentage points.[24] We also separately estimate the effect on log-loans from different patron groups: those who have a Harvard affiliation and those who do not. The effect is strongest among people who are not directly affiliated with Harvard. Assuming that checking out a book at Harvard involves a bigger hassle for these consumers, this suggests that the substitution effect of digitization is largest when traditional means of obtaining a book are the most costly. Note that search costs are not the only difference between the loans sample and the sales sample since Harvard patrons might differ from general consumers on other dimensions. Yet, we consider these results as in line with our hypothesized mechanism.

We also explore a setting in which the substitution mechanism is muted: books that are digitized and hence searchable, but not fully made available to consumers. Ideally, we would apply the same estimation strategy as above, comparing sales of books that are never digitized with books that are made partially available before and after their digitization. Unfortunately, while the Google Books project provided such partial digitization for in-copyright books, Harvard did not participate in this digitization, and we therefore do not observe the dates of digitization for these titles. Instead, we explore a much more modern sample of books that are available on Amazon, and we focus on Amazon's "Search Inside the Book" (SITB) feature. Amazon digitizes a subset of its books and allows consumers to search their texts, but it only provides snippets of the full text to look at before purchase. We compare the current ranks and total Amazon ratings of books that have the SITB feature with books *by the same author* without the feature, across a total of 11,166 recent books.[25] We find significant differences in our demand measures: Books that are included in

[24] We report results from robustness checks in appendix Table D.5.

[25] We provide more detail on the program and how it helps us identify the discovery function of digitization in appendix Section B, and we summarize our results in appendix Table D.6.

the SITB program have significantly higher demand than those that were not digitized at all. In particular, conditional on the author and publication year, books with the SITB feature have received over 200 more reviews, and are ranked over 50 percent more highly. Because authors and publishers can choose whether to include their books in the SITB program, this comparison does not imply a causal relationship. Still, it provides suggestive evidence that the search function can increase sales through other channels.

4.4 Digitization and the Supply Side

In addition to providing an opportunity for consumers to learn about products they wouldn't otherwise be aware of, it is possible that digitization of these lesser-known and perhaps forgotten works affects the supply of physical editions. Google Books may enable publishers to create and publish more and higher-quality copies of public domain books. We therefore examine the effect of digitization on the supply of the digitized products, including the number of available editions and their prices, as well as whether the new editions can explain the digitization-related increase in sales.

Table 6 presents the regression results from the respective variants of equation (1). The first four columns report the estimated effects of digitization on the number of editions. Overall, we find that the digitization of the public domain works through Google Books had a statistically significant effect of about two more new editions per book and year (column 1). While major publishers likely already had the means to obtain these texts, independent publishers may not have had the same resources and therefore may have benefited much more from the digitization project. Consistent with this, we find that the increase in editions is driven almost entirely by independent publishers (columns 2 and 3). Further, while one might expect the independent publishers to produce lower-quality and hence cheaper editions, we find no discernible difference in prices across the new and previous editions.[26] Naturally, the increases in the number of *new* editions imply that the number of *total* editions increases as well (column 4).

Columns 5 and 6 of Table 6 turn to the question of how the supply of editions influences downstream readership. In column 5, we estimate the causal relationship between available editions and log-sales in a two-stage least squares regression. In the first stage of this regression, we instrument for the cumulative number of available editions with the book's digitization status. That is, column 4 also functions as the first stage to the column 5 regression. We find that the number of copies sold increases statistically significantly with each additional edition, but only by about 0.9 percent. Column 6 returns to OLS regressions. We utilize the exogenous timing of digitization to identify the effects of both digitization and the number of available

[26] See appendix Section C and appendix Table D.7.

editions on downstream demand. We find that the positive effect of digitization on log-sales remains significant but decreased by about half, to 2.6 percent. This suggests that about half of the positive effect of digital provision on sales can be explained by changes on the supply side. The remainder of the effect seems to be due to an improved information environment for consumers.

5 Discussion

Our empirical analysis shows that free digital distribution may increase rather than decrease the sales of physical works. This result is particularly striking given that we study books that are provided in their entirety for digital consumption and where the potential for cannibalization is high. The positive effect of free digital provision on demand is stronger for more obscure books and spills over to a digitized author's non-digitized books. Our results are consistent with the idea that digitization allows readers to discover new works, some of whom subsequently choose to consume the physical product. We also find that digital distribution encourages the publication of new physical editions, suggesting that digital distribution can stimulate the supply of physical products as well. Overall, our results contribute to helping uncover the potential for digital distribution to stimulate demand for and supply of physical products.

Beyond our theoretical contributions, our findings inform how changes in representation and aggregation of works shape management strategy in the publishing industry. Publishers and authors had a mixed reaction to the arrival of mass digitization projects with many being opposed due to concerns around cannibalization. Our work shows that publishers and authors might want to actively distribute digital versions of their works to boost physical sales. While our work shows that even full text access can stimulate demand, for in-copyright works, publishers might want to consider "snippet" access, which allows for the full power of text-based discovery but mutes incentives for cannibalization. In such contexts, the positive impacts could be even stronger. In fact, our results are aligned with an internal study conducted by Amazon that found a 9% increase in sales for books that had the SITB feature enabled.[27] Further, even if publishers are wary of digital distribution in general, they might consider a selective strategy where less popular books, which are at a lower risk of cannibalization, are provided in digital form, potentially even for free.

Our results also have implications for ongoing legal and policy debates on the design of copyright law for the digital age. First, we provide causal, empirical evidence for the theoretical debate about whether free digital distribution cannibalizes sales or promotes discovery. By some calculations, about 12,686 copyrighted books are available to be digitized for every year between 1923-1936 (Reimers, 2019). Not only

[27] https://bit.ly/3rEJupW

could these books be made available for digital access, but digitization might also increase the sales of their physical editions. Therefore, our results help strengthen the value proposition of mass-digitization projects such as Google Books, Project Gutenberg or the Hathi Trust. Paving the way for further investment in such projects might translate to easing access to past knowledge and unlocking large benefits in terms of follow-on innovation and creativity (Biasi and Moser, 2018; Furman et al., 2018).

Our work is not without limitations. First, our intention is not to say that search-enabled digital distribution will necessarily increase physical sales. Rather, our goal is more modest: to establish that such positive effects are possible. Whether or not digital distribution enabled by search increases sales depends on contextual factors shaping the relative balance between the forces of cannibalization and discovery. In our setting, discovery is made possible by Google Books' full text search feature and the project also offers a relatively poor substitute for a physical book – reading the entire text of a book on the website is not convenient. In other contexts, the net effect might depend on the quality of the digital substitute and the availability and capacity of the search technology to drive discovery. Examining these conditions in more detail offers an exciting avenue for future work. Second, even though we provide some evidence to suggest that our results might generalize to in-copyright works, we study full-text access for out-of-copyright works. Future work should examine the effects of snippet access to in-copyright works on offline demand more directly. Finally, the effects of digital distribution on physical supply also need further investigation. In particular, since publishers do not need to negotiate licenses to republish public domain works, our effects are likely an upper bound on the potential effects of digital provision on online supply and need further scrutiny.

In sum, digital distribution and search offers a powerful tool to reshape how content is produced, discovered, consumed and distributed. Deepening our understandind of how consumers discover and consume content in a world with both physical and digital channels remains an exciting topic for future research.

References

Aguiar, L. and J. Waldfogel (2018, March). As streaming reaches flood stage, does it stimulate or depress music sales? *International Journal of Industrial Organization 57*, 278–307.

Bai, J. and J. Waldfogel (2012, December). Movie piracy and sales displacement in two samples of Chinese consumers. *Information Economics and Policy 24*(3-4), 187–196.

Baye, M. R., B. D. l. Santos, and M. R. Wildenbeest (2015). Searching for physical and digital media: The evolution of platforms for finding books. In *Economic Analysis of the Digital Economy*, pp. 137–165. University of Chicago Press.

Berger, J., A. T. Sorensen, and S. J. Rasmussen (2010). Positive effects of negative publicity: When negative reviews increase sales. *Marketing science 29*(5), 815–827. Publisher: INFORMS.

Biasi, B. and P. Moser (2018). Effects of copyrights on science-evidence from the US book republication program. Technical report, National Bureau of Economic Research.

Bounie, D., M. Bourreau, and P. Waelbroeck (2006). Piracy and the Demand for Films: Analysis of Piracy Behavior in French Universities. *Review of Economic Research on Copyright Issues 3*(2), 15–27.

Brynjolfsson, E., Y. J. Hu, and M. D. Smith (2003, November). Consumer Surplus in the Digital Economy: Estimating the Value of Increased Product Variety at Online Booksellers. *Management Science 49*(11), 1580–1596. Publisher: INFORMS.

Brynjolfsson, E., Y. J. Hu, and M. D. Smith (2006). From niches to riches: Anatomy of the long tail. *Sloan Management Review 47*(4), 67–71.

Chen, H., Y. J. Hu, and M. D. Smith (2019, January). The Impact of E-book Distribution on Print Sales: Analysis of a Natural Experiment. *Management Science 65*(1), 19–31.

Chiou, L. and C. Tucker (2017). Content aggregation by platforms: The case of the news media. *Journal of Economics & Management Strategy 26*(4), 782–805. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/jems.12207.

Danaher, B., S. Dhanasobhon, M. D. Smith, and R. Telang (2010, November). Converting Pirates Without Cannibalizing Purchasers: The Impact of Digital Distribution on Physical Sales and Internet Piracy. *Marketing Science 29*(6), 1138–1151.

Danaher, B., M. D. Smith, and R. Telang (2014, January). Piracy and Copyright Enforcement Mechanisms. *Innovation Policy and the Economy 14*, 25–61.

Ellison, G. and S. F. Ellison (2018, January). Match Quality, Search, and the Internet Market for Used Books. Technical Report w24197, National Bureau of Economic Research.

Forman, C., A. Ghose, and A. Goldfarb (2009). Competition between local and electronic markets: How the benefit of buying online depends on where you live. *Management science 55*(1), 47–57. Publisher: INFORMS.

Furman, J. L., M. Nagler, and M. Watzinger (2018). Disclosure and subsequent innovation: Evidence from the patent depository library program. Technical report, National Bureau of Economic Research.

Ghose, A., M. D. Smith, and R. Telang (2006, March). Internet Exchanges for Used Books: An Empirical Analysis of Product Cannibalization and Welfare Impact. *Information Systems Research 17*(1), 3–19.

Google (2007). World's oldest publisher stays at the cutting edge with Google Book Search.

Greenstein, S., J. Lerner, and S. Stern (2013). Digitization, innovation, and copyright: What is the agenda? *Strategic Organization 11*(1), 110–121. Publisher: Sage Publications Sage UK: London, England.

Heald, P. J. (2007). Property rights and the efficient exploitation of copyrighted works: an empirical analysis of public domain and copyrighted fiction best sellers. *UGA Legal Studies Research Paper* (07-003).

Holtz, D., B. Carterette, P. Chandar, Z. Nazari, H. Cramer, and S. Aral (2020). The Engagement-Diversity Connection: Evidence from a Field Experiment on Spotify. In *Proceedings of the 21st ACM Conference on Economics and Computation*, pp. 75–76.

Kumar, A., M. D. Smith, and R. Telang (2014). Information discovery and the long tail of motion picture content. *Mis Quarterly 38*(4), 1057–1078. Publisher: JSTOR.

Nagaraj, A. (2018). Does Copyright Affect Reuse? Evidence from the Google Books Digitization Project. *Management Science*.

Nagaraj, A. (2020). The Private Impact of Public Data: Landsat Satellite Maps and Gold Exploration. *March 5*(2020), 5.

Ni, J., J. Li, and J. McAuley (2019). Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 188–197.

Oberholzer-Gee, F. and K. Strumpf (2010, January). File Sharing and Copyright. *Innovation Policy and the Economy 10*, 19–55.

Reimers, I. (2019). Copyright and generic entry in book publishing. *American Economic Journal: Microeconomics 11*(3), 257–84.

Rob, R. and J. Waldfogel (2006, April). Piracy on the High C's: Music Downloading, Sales Displacement, and Social Welfare in a Sample of College Students. *The Journal of Law and Economics 49*(1), 29–62.

Rob, R. and J. Waldfogel (2007). Piracy on the Silver Screen. *The Journal of Industrial Economics 55*(3), 379–395.

Samuelson, P. (2009, July). Legally speakingThe dead souls of the Google book search settlement. *Communications of the ACM 52*(7), 28.

Sismeiro, C. and A. Mahmood (2018). Competitive vs. complementary effects in online social networks and news consumption: A natural experiment. *Management Science 64*(11), 5014–5037. Publisher: INFORMS.

Smith, M. D. and A. Zentner (2016, January). Internet effects on retail markets. *Handbook on the Economics of Retailing and Distribution*. ISBN: 9781783477388 Publisher: Edward Elgar Publishing Section: Handbook on the Economics of Retailing and Distribution.

Somers, J. (2017, April). Torching the Modern-Day Library of Alexandria. *The Atlantic*.

Waldfogel, J. (2017). How digitization has created a golden age of music, movies, books, and television. *Journal of economic perspectives 31*(3), 195–214.

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

Watson, J. (2017). What is the Value of Re-use? Complementarities in Popular Music.

Yu, Y., H. Chen, C.-H. Peng, and P. Y. Chau (2018). The Causal Effect of Subscription Video Streaming on DVD Sales: Evidence from a Natural Experiment. *Available at SSRN 2897950*.

Zentner, A. (2005, October). File Sharing and International Sales of Copyrighted Music: An Empirical Analysis with a Panel of Countries. *The B.E. Journal of Economic Analysis & Policy 5*(1).

Zentner, A. (2006, April). Measuring the Effect of File Sharing on Music Purchases. *The Journal of Law and Economics 49*(1), 63–90.

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

6 Tables and Figures

Tables

Table 1. **Summary Statistics**

Panel A: Book-Level

	N	Mean	Std. Dev.	Median	Min	Max
Scanned (0/1)	88006	0.43	0.49	0	0	1
Year Scanned	37717	2006.98	1.19	2007	2005	2009
Total Loans (2003-11)	88006	2.23	5.33	1	1	1130
Total Sales (2003-11)	9204	4990.54	56486.76	0	0	1965285
Total Editions (2003-11)	88006	3.21	14.85	0	0	842
Popular (0/1/2)	9204	0.13	0.46	0	0	2

Panel B: Book-Year Level

	N	Mean	Std. Dev.	Median	Min	Max
Post-Scanned (0/1)	792054	0.19	0	0	0	1
Loans	792054	0.25	1	0	0	189
Sales	82836	554.50	6839	0	0	626610
Any-Loans (0/1)	792054	0.17	0	0	0	1
Any-Sales (0/1)	82836	0.16	0	0	0	1
Annual Editions	792054	0.36	3	0	0	542

Note: This table lists summary statistics for the full sample. Observations in Panel A are at the book-level for 88,006 books in the main sample with at least one loan over the study period. Observations in Panel B are at the book-year level for a balanced panel of 792,054 observations (88,006 books over 9 years from 2003 to 2011). Scanned: 0/1 for books that have been digitized in the time period 2003 to 2011. 37,743 books were digitized by the Google Books project and statistics for the Year Scanned variable are calculated from this subset. Sales data were collected for a subset of 9,204 books and summary statistics are from this subgroup. Popular = 1 for books that had more than one loan before the digitization program started (i.e., in 2003 and 2004). Any Loans and Any Sales are indicators = 1 if a book was loaned or sold at least once in a given year. See text for more details.

Table 2. **Baseline Estimates for the Impact on Sales**

	Log-OLS		LPM	
	(1) Log-Sales	(2) Log-Sales	(3) Any-Sales	(4) Any-Sales
Post-Scanned	0.0480***	0.0466***	0.0782***	0.0770***
	(0.0125)	(0.0130)	(0.00481)	(0.00487)
Book FE	Yes	Yes	Yes	Yes
Year FE	Yes	No	Yes	No
Year-location FE	No	Yes	No	Yes
N	82836	82836	82836	82836

Note: This table presents estimates from OLS models evaluating the overall impacts of book digitization on sales. Columns (1) and (2) report results from log-OLS models, where the dependent variable is $\ln(sales)$, and columns (3) and (4) report results from linear probability models (LPMs), where the dependent variable is an indicator that is 1 if book j had at least one sale in year t. Post-Scanned equals one in all years after a book has been digitized. Popular equals one for books that had more than one loan before the digitization program started (i.e., in 2003 and 2004). All models include book and year fixed effects. Columns (2) and (4) interact the year fixed effects with library-location fixed effects. Standard errors are in parentheses, clustered at the book level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 3. **Robustness Checks**

	Endogeneity			Sample		Model		
	(1) Controls	(2) Pusey/2005	(3) Pusey/2005	(4) Public Domain	(5) Twins	(6) Asin(sales)	(7) Poisson	(8) Logit (0/1)
Post-Scanned	0.0438***	0.0496***	0.0488***	0.0508***	0.0614***	0.0676***	0.297*	0.647***
	(0.0129)	(0.0161)	(0.0160)	(0.0140)	(0.0169)	(0.0151)	(0.153)	(0.0868)
Book FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	No	No	No	No	No	No	Yes	Yes
Year-location FE	Yes	Yes	Yes	Yes	Yes	Yes	No	No
Addtl. Controls	Yes	No	Yes	No	No	No	No	No
N	82836	62199	62199	29394	36738	82836	26586	22671

Note: This table evaluates robustness of the baseline regressions to alternate specifications and sample restrictions. Columns (1) through (5) provide zero-inflated Log-OLS estimates (i.e., the dependent variable is $Ln(Loans_{it}+1)$ or $Ln(Sales_{it}+1)$). Column (1) adds controls for a book's Google Search volume as well as a dummy variable that equals one if the book has also been digitized on Project Gutenberg before year t. Column (2) drops all books digitized in 2005 or located in Pusey 3, and column (3) adds the demand controls from (1) to the sample from (2). Columns (4) and (5) introduce alternate sample restrictions, limiting the sample to books that are in the public domain and therefore digitized at some point in our sample period (4), or including only matched pairs (digitized and not) that are located exactly next to each other (5). The remaining columns vary the functional form. Column (6) uses the hyperbolic sine of sales as the dependent variable, column (7) provides estimates from a Poisson regression, and column (8) estimates the likelihood that a book is sold at all in year t in a binomial logit regression. Post-scanned equals one in all years after the book has been digitized. All models include book and year fixed effects. Columns (1) through (6) additionally interact these year fixed effects with library-location fixed effects. Standard errors are in parentheses, clustered at the book level. * $p<0.10$, ** $p<0.05$, *** $p<0.01$

Table 4. **Exploring the Discovery Mechanism**

Panel A: Effects by the Book's Popularity

	Baseline		Robustness	
	(1) Log-Sales	(2) Log-Sales	(3) Public Domain	(4) Twins
Post-scanned:				
... × Presales=0	0.0456***	0.0453***	0.0492***	0.0575***
	(0.0123)	(0.0128)	(0.0141)	(0.0165)
... × Presales>0	0.324***	0.309***	0.319***	0.364**
	(0.103)	(0.104)	(0.103)	(0.142)
... × Presales>500	-0.112	-0.133	-0.125	-0.0860
	(0.101)	(0.100)	(0.100)	(0.149)
Book FE	Yes	Yes	Yes	Yes
Year FE	Yes	No	No	No
Year-location FE	No	Yes	Yes	Yes
N	82836	82836	29394	36738

Panel B: Spillovers to the Author's Other Books

	In-Copyright		All books	
	(1) Log-Sales	(2) Any-Sales	(3) Log-Sales	(4) Any-Sales
Post-scanned ×				
... this book			0.0553*	0.0878***
			(0.0316)	(0.00959)
... other book	0.103**	0.0346***	0.0663*	0.0299***
	(0.0460)	(0.0108)	(0.0354)	(0.00891)
Book FE	Yes	Yes	Yes	Yes
Year-location FE	Yes	Yes	Yes	Yes
N	19224	19224	29799	29799

Note: Panel A reports the heterogeneous impact of book digitization on sales for books of varying popularity. Presales=0 includes books with no sales in 2003 and 2004. Presales>0 describes books with 1 to 500 total sales in 2003 and 2004. And Presales>500 includes all books with more than 500 sales. All columns report results from zero-inflated Log-OLS regressions. Columns (1) and (2) mirror the first two baseline models in Table 2. Columns (3) and (4) repeat the sample-related robustness checks from Table 3, using only public domain (digitized) books (3) and including only matched pairs (digitized and not) that are located exactly next to each other (4). Panel B evaluates the impact of digitization on sales of the digitized book (this book) as well as on the sales of other books by the same author (other book). The first two columns use only books that are not digitized at all, whereas the last two columns also include digitized books. All estimations are run on authors who have at least two books in our sample. Columns (1) and (3) show results from zero-inflated log-OLS regressions, and columns (2) and (4) show results from linear probability models. Column (1) in Panel A uses book and year fixed effects. All other models include book and year-location fixed effects. Standard errors are in parentheses, clustered at the book level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 5. **Effects on Harvard Library Loans**

	Main effect		Consumer groups	
	log-OLS	LPM	Non-Harvard	Harvard
Post-Scanned	-0.0511***	-0.0613***	-0.0362***	-0.0157***
	(0.00152)	(0.00170)	(0.00121)	(0.000964)
Book FE	Yes	Yes	Yes	Yes
Year-Location FE	Yes	Yes	Yes	Yes
N	792054	792054	792054	792054

Note: This table reports effects of digitization on loans at Harvard's libraries. Columns (1) and (2) provide baseline estimates from a zero-inflated Log-OLS estimation (1) and a linear probability model (2). Columns (3) and (4) show heterogeneous effects based on our popularity measures from Panel A, for a log-OLS model and a linear probability model, respectively. Finally, columns (5) and (6) disambiguate the effects from the log-OLS model for loans from non-Harvard affiliated consumers (5) and Harvard affiliated consumers (6). All models include book and year-location fixed effects. Standard errors are in parentheses, clustered at the book level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

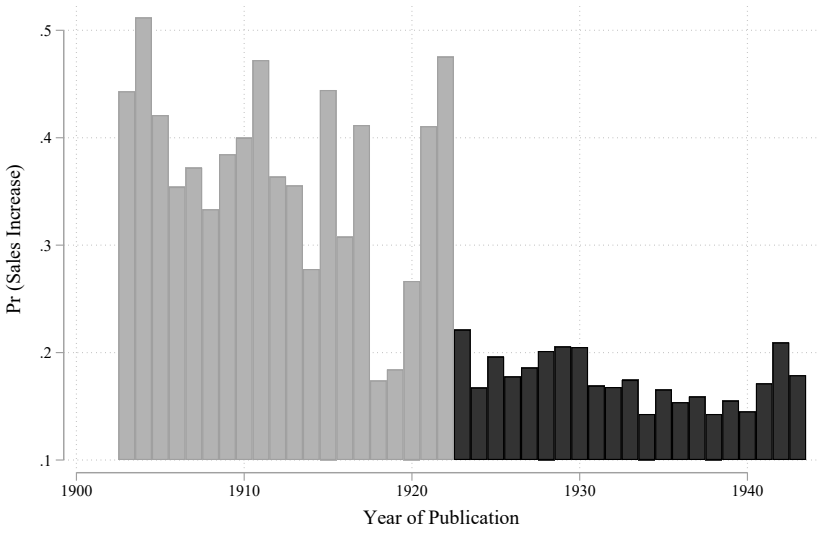
Table 6. **The Role of New Editions**

	Effect on Editions				Effect on Sales	
	(1) New Eds.	(2) Majors	(3) Indies	(4) Cumulative	(5) Log-sales	(6) Log-sales
Post-Scanned	2.091***	0.0302***	2.061***	5.061***		0.0259**
	(0.130)	(0.00581)	(0.127)	(0.328)		(0.0132)
Cumulative editions					0.00920***	0.00408***
					(0.00166)	(0.000635)
Book FE	Yes	Yes	Yes	Yes	Yes	Yes
Year-location FE	Yes	Yes	Yes	Yes	Yes	Yes
N	82836	82836	82836	82836	82836	82836

Note: This table investigates the role of new edition releases. Columns (1) through (4) examine the effect of digitization on the number of editions. The dependent variables are the total number of new editions of that title (1), the number of new editions of the title released by the five major publishers (2), the number of new editions by non-major publishers (3), and the cumulative number of available editions of the book (4). Column (4) also serves as the first stage to column (5), which reports the second stage results of an instrumental variables estimation of the impact of availability (cumulative editions) on zero-inflated log-sales. Column (6) returns to a regular OLS regression of zero-inflated log-sales as a function of digitization and availability. Post-scanned equals one in all years after the book has been digitized. All models include book and year-location fixed effects. Standard errors are in parentheses, clustered at the book level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

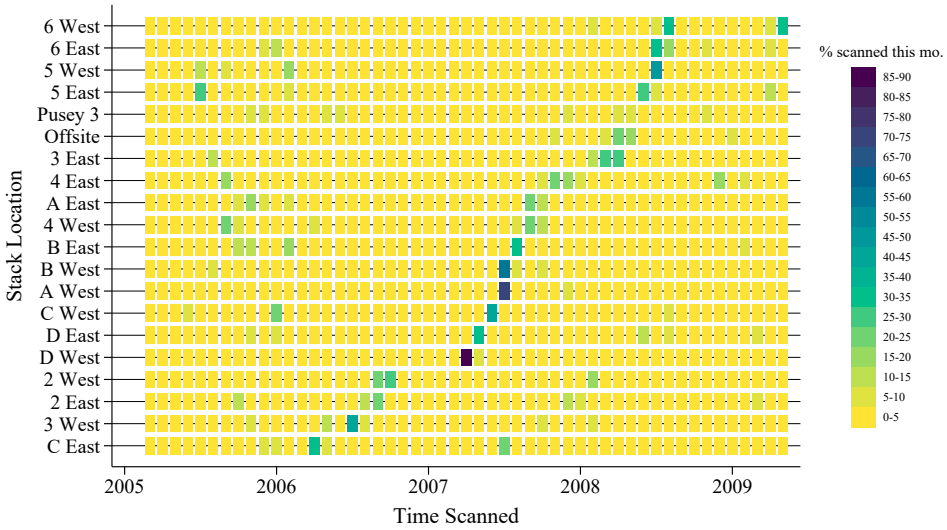
Figures

Figure 1. **Comparing Change in Demand for Pre-1923 and Post-1923 Books**



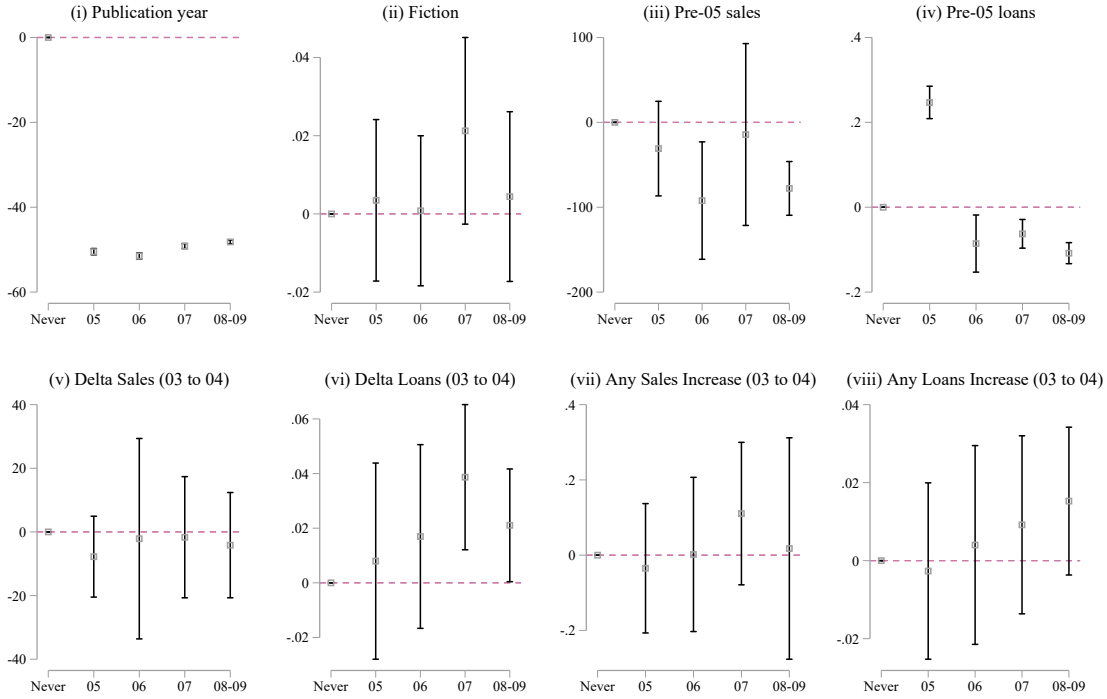
Note: This figure explores the impact of the digitization program on a cross-sectional sample of English language books originally published between 1904-1942, of which only those published before 1923 are digitized due to copyright restrictions. This includes 6,755 books with sales data. For each book, we calculate the change in the number of sales in the 2010-2011 period (after digitization) as compared to the 2003-2004 period (before digitization). We then plot the share of books in each publication year that increase their sales on the y-axis and the publication year on the x-axis. Books published after 1923 (which were not scanned) are indicated in black, and those before are indicated in gray.

Figure 2. **Timing of Book Digitization By Library Shelf Location**



Note: This figure provides an illustration of the timing of book digitization by shelf location for 30,839 (of 37,317) scanned books for which we know the exact shelf location in Harvard's Widener library. For each shelf location, we calculate the percent of books digitized in a given calendar month; the bluer the rectangle the higher percent of books from that location are digitized in that month. Shelves are sorted in ascending order (from bottom to top) of the month in which the maximum percent of their books were digitized. See text for more details.

Figure 3. **Comparing Books by Scanned Year**

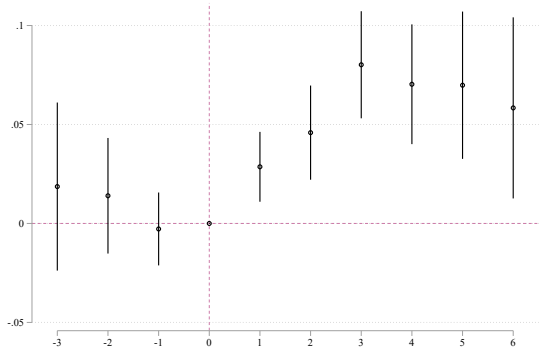


Note: This figure compares the full sample of books depending on the year in which they were scanned with unscanned books. Each panel presents coefficients on the years of digitization from title-level cross-sectional regressions of different book-level covariates on subject dummies, library location dummies, and year-of-digitization dummies, for all books in the dataset. The dependent variables in the first row are (i) year of publication, (ii) a dummy for whether a book is fiction, (iii) pre-05 sales and (iv) pre-05 loans. Variables in the second row provide a measure of trends pre-digitization between 2003 and 2004 recording (v) change in sales, (vi) change in loans, and an indicator for an increase in (vii) sales and (viii) loans. We plot coefficients for each year of digitization, including 95% confidence intervals (using robust standard errors). The omitted category is books that were not scanned.

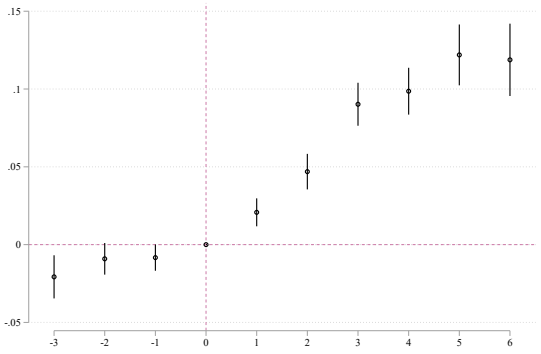
Figure 4. **Time-Varying Estimates of the Impact of Digitization**

Panel A: Full Sample

(i) Log-Sales

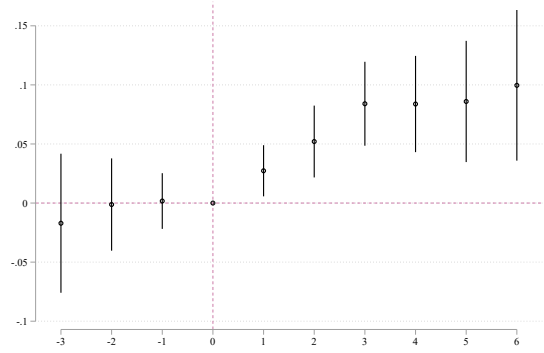


(ii) Any-Sales

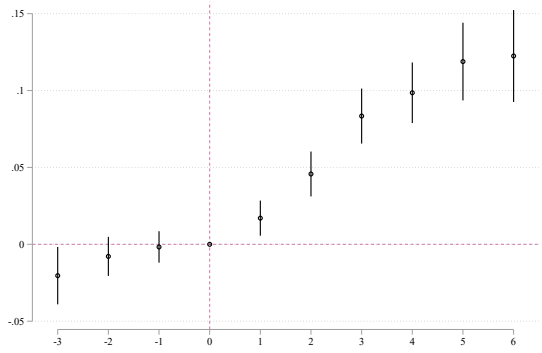


Panel B: Twins Sample

(i) Log-Sales



(ii) Any-Sales



Note: This figure provides visual illustrations of the event study specification: $Y_{it} = \alpha + \Sigma_z \, \beta_z (scanned)_i \times 1(z) + \gamma_i + \mu_{it} + \varepsilon_{it}$, where γ_i and μ_t represent book and year-location fixed effects, respectively, for book i and year t, $(scanned)_i$ equals one for all books that were eventually scanned and z represents the "lag," or the number of years that have elapsed since a book was first digitized ($= 0$ in the year before digitization). The main dependent variables are Log-Sales (panels (*i*)) or Any-Sales (panels (*ii*)). Panel A uses the full sample of all 9,024 titles with sales information. Panel B includes only matched pairs (digitized and not) that are located exactly next to each other, for a total of 4,082 titles. The chart plots values of β_z for different values of z.

Online Appendix

A Regression Discontinuity

The main analysis takes advantage of variation in the timing and status of digitization across all books in the Harvard Widener library system. An underlying assumption in these analyses is that books that are digitized are inherently similar to books that are not, or not yet, digitized. However, whether a book is digitized at all is a function of the book's copyright status. Throughout the time period of our study, all works that were originally published before 1923 are in the public domain and hence digitized, whereas works from 1923 and later were still protected by copyright and hence not digitized.

This discontinuity in copyright status is due to the most recent copyright extension in the United States. The 1998 Copyright Term Extension Act retroactively extended the copyright term for all protected works by twenty years, from 75 to 95 years for the works in our dataset. This extension provides a sharp, exogenous discontinuity in the ex-post digitization status for works originally published around 1923. Beyond their digitization status, however, nothing changed systematically for just one group of titles over the time period in our study. Thus, were it not for the digitization of the older works, one might expect analog demand for these titles to evolve similarly for works originally published on both sides of the 1923 cutoff.

We therefore examine how demand changed from 2003/04 (before any works were digitized) to 2010/11 (after digitization was completed) in a regression discontinuity design that formalizes the patterns shown in Figure 1 in the main text. Formally, we utilize the jump in digitization around the original publication year 1923 to estimate regression equations of the form

$$Y_i = \alpha + \beta Digitized_i + k(year_i) + \varepsilon_i, \tag{3}$$

where Y_i describes various measures of the change in book j's unit sales from 2003/04 to 2010/11, including the absolute unit changes, an asymptotic sine transformation of these changes, and an indicator that equals one if there is an increase in book i's demand.[28] Moreover, $Digitized_i$ is an indicator variable that is 1 if the book was digitized, which is a deterministic function of the book's original year of publication. We define $k(year_i)$ as a quadratic function of the book's publication year, centered around 1923. The bandwidth in each specification is its mean-squared-error optimal bandwidth.

Table D.1 shows the results from these specifications. All results support those in the main text: treat-

[28] We use an asymptotic sine transformation instead of the more common log transformation because one would naturally expect many negative changes in demand, and dropping these may bias results (note that our dependent variable is the *change* in demand).

ment through digitization leads to an increase in marketwide sales. The estimated percentage effects of around 28 percent are even larger than those from our main specifications. These results are robust to different bandwidths and functional forms of the publication year. Figure D.2 further plots the annual coefficients of the regressions on the likelihood of increased demand (analog to column 6 in Table D.1). It illustrates that books originally published before 1923 and therefore digitized by Google Books are significantly more likely to experience an increase in sales.

B Amazon Appendix

In this section, we present additional results that explore (a) the impact of digitization on demand for scanned books as measured by the number of Amazon reviews and (b) the impact of the Amazon "Search Inside The Book" (SITB) feature on book demand on Amazon.

B.1 The Effect of Google Books Digitization on Amazon Reviews

Our baseline analysis examines the effect of Google Books digitization of Harvard library books on retail sales. While these sales data do include sales from online channels like Amazon, we directly examine the impact of book digitization on demand for books on Amazon as an additional robustness check. Lacking panel data on sales volume, we examine the effect on the annual quantity of reviews, with the assumption that books with greater sales also have a higher number of reviews.

Specifically, we try to match titles of books in our Harvard sample of public domain books with titles of all books on Amazon provided by Ni et al. (2019) and available on https://nijianmo.github.io/amazon/index.html. We restrict our attention to public domain books since these are more likely to be available on Amazon, improving our match rate. We find matches for 559 scanned books in our sales sample. Reviews end in 2018 and go as far back as 1997. We construct a balanced sample for our books between the years 1997-2018 for a total of 12,298 observations. On overage books have about 1.07 reviews per year. We then estimate a version of our baseline regression from equation (1), where Y_{it} represents the number of annual Amazon reviews. Estimates from this regression are presented in columns 1 and 2 of Table D.6. As is clear from these results, book digitization by Google Books significantly increases the number of reviews for scanned books. The estimates imply a fourfold increase in the number of annual reviews on Amazon, albeit against a low base of 1.07 reviews. These results suggest that the baseline effect on increased sales we document is likely to translate to an increase in consumer reviews on Amazon as well.

B.2 Examining the Effect of Digitization on Modern Sample of In-Copyright Books

Our main analysis has two major advantages: it relies on a clean natural experiment, and it allows us to quantify the effect of the full provision of digital texts on analog demand. Along with these advantages come some disadvantages. First, we only observe digitization of books that are in the public domain and therefore quite old. And second, the full-text digitization prevents the isolation of the discovery effect by muting the substitution effect. We address both concerns in a parallel analysis that focuses on Amazon's "Search Inside the Book" (SITB) program, which scans the entire text of books and makes them available for full-text search. However, the book itself is provided only in "snippet" view, so that the SITB feature is unlikely to displace regular sales. If digitization-enabled search helps consumers discover new content, then books that adopt SITB are likely to have greater demand than books that do not adopt SITB.

Unlike Google Books, which was implemented en masse, the SITB feature is optional for publishers and authors to adopt.[29] An ideal experiment would randomly assign SITB status to a set of comparable books and examine the effect of this treatment on sales. Lacking such an experiment, we provide evidence from a cross-sectional fixed-effects specification for the hypothesis that SITB adoption raises sales. We start with a set of the top 1200 featured books in the "Business and Investing" and "Education and Reference" categories on Amazon. For these 2400 books, we sample the set of authors and obtain all other books they have written. We thus build a sample of 1175 authors who have written a total of 11,166 books. Of these books, about 80 percent have the SITB feature enabled. While some of this variation is due to time (more recent books are more likely to have SITB enabled), there is considerable variation within years and authors.

Using this sample, we estimate the following specification: $Y_{ia} = \alpha + \beta SITB_{ia} + \gamma_a + \delta_i + \varepsilon_{ia}$, where Y_{ia} indicates the total number of reviews or the natural log of the current bestseller rank for book i published by author a. $SITB_{ia}$ equals one if the book has SITB enabled, and zero otherwise. γ_a indicates author fixed effects and δ_i indicates publication year fixed effects. The key identification concern with this specification is that books with greater market potential are selectively SITB enabled, while more marginal books are not. Controlling for author and year fixed effects helps considerably as we are not comparing higher selling authors vs. other authors, or years with greater demand vs. years with lower demand.

Estimates from this analysis are presented in Table D.6. The coefficients are positive and significant. As against a base of 554 reviews, SITB enabled books see about 214 more reviews (Col 2), an increase of about 38 percent. Similarly, SITB enabled books have about a 73 percent lower (better) rank compared to books that are not part of the SITB program. Taken at face value, these results suggest that even for a sample of

[29] This implies that the program also includes more recent books.

modern, in-copyright books, book digitization can significantly improve discovery of new content, thereby increasing demand. Note however that we cannot rule out the concern that authors selectively treat their better books with SITB. Our results in this section should therefore be seen as preliminary and worthy of future research, rather than conclusive.

C The Impact on Prices

The main text shows that digitization through Google Books leads to an increase in the number of editions, particularly from independent publishers. The digitization-induced increase in unit sales could be driven by a decrease in prices of the new editions. If such decreases in prices are large enough, then digitization could lead to decreases in revenues despite the increase in sales. We therefore examine the effect of digitization on the prices of new editions.

While we do not have price information in the sales sample, we obtain list prices from Bowker's Books-in-Print sample. We treat each newly published edition as an observation, and we estimate the edition's suggested retail price as a function of the title's digitization status at the time of the edition's publication. We further include indicators for the edition's year of publication and for each title. Formally, we estimate

$$Y_{jit} = \alpha + \beta PostScanned_{jit} + \gamma_i + \lambda_t + \varepsilon_{jit}, \tag{4}$$

where Y_{jit} is the suggested retail price (or its log) of edition j of title i, which was published in year t. In addition, $PostScanned_{jit}$ is an indicator that equals 1 if title i was digitized before the year t in which edition j was published, and γ_i and λ_t are title and publication year indicators, respectively. In additional checks, we add controls for the number of available editions for title i.

Table D.7 depicts the results from these regressions. None of the specifications show any evidence that Google's digitization program had an impact on prices of new editions, as all coefficients are small and statistically insignificant.

D Appendix Tables and Figures

Table D.1. **Regression Discontinuity Estimates**

	Sales		
	(1) sales	(2) asinh(sales)	(3) increase
Digitized	577.4	0.277**	0.159***
	(450.8)	(0.126)	(0.0229)
N	8016	8016	8016

Note: This table presents results from regression discontinuity estimations on the title level. The dependent variables are functions of the changes in analog sales between 2003/04 (before digitization) and 2010/11 (after digitization). In column 1, it is the absolute change in analog sales; column 2 uses the asymptotic sine of that change; and column 3 uses an indicator that is 1 if analog demand has increased. The independent variable of interest, *Digitized*, is an indicator that is 1 if the book was digitized (i.e. originally published before 1923). A quadratic function of the publication year is included. The bandwidth in each specification is the MSE-optimal bandwidth. Robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table D.2. **Varying Effects by Scan Year**

Panel A: Effects on Log-sales (OLS)

	Individual Scan Year				
	2005	2006	2007	2008	2009
Post-Scanned	0.0198	0.0751***	0.0289	0.0531**	0.171**
	(0.0263)	(0.0243)	(0.0219)	(0.0226)	(0.0747)
Book FE	Yes	Yes	Yes	Yes	Yes
Year-Location FE	Yes	Yes	Yes	Yes	Yes
N	60480	60345	61353	60201	54261

Panel B: Effects on Any-sales (LPM)

	Individual Scan Year				
	2005	2006	2007	2008	2009
Post-Scanned	0.0675***	0.0875***	0.0804***	0.104***	0.154***
	(0.00907)	(0.00995)	(0.00891)	(0.0103)	(0.0372)
Book FE	Yes	Yes	Yes	Yes	Yes
Year-Location FE	Yes	Yes	Yes	Yes	Yes
N	60480	60345	61353	60201	54261

Note: This table shows the effect of digitization on sales across digitization cohorts. Panel A reports results from OLS estimations where the dependent variable is zero-inflated log-sales; Panel B reports estimates from linear probability models where the dependent variable is an indicator that is one if at least one copy of the book was sold. In both panels, the first five columns report results from separate regressions for each digitization cohort. For example, in column (1) we keep all books scanned in 2005 as well as all unscanned books, and we estimate the effect of digitization in that year. We do the analogous exercise for subsequent scan years in columns (2) through (5). Post-scanned equals one in all years after the title has been digitized. All models include book and year-location fixed effects, and standard errors (in parentheses) are clustered at the book level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table D.3. **Robustness Checks for Sales, by Popularity**

	Endogeneity			Model		
	(1) Controls	(2) Pusey/2005	(3) Pusey/2005	(4) Asin(sales)	(5) Poisson	(6) Logit
Post-scanned:						
... × Presales=0	0.0458***	0.0440***	0.0457***	0.0669***	0.999***	0.780***
	(0.0127)	(0.0159)	(0.0158)	(0.0150)	(0.374)	(0.0895)
... × Presales>0	0.279***	0.369***	0.341**	0.350***	0.289***	-3.634***
	(0.101)	(0.139)	(0.135)	(0.115)	(0.110)	(0.369)
... × Presales>500	-0.222**	-0.0570	-0.125	-0.148	0.293*	9.060
	(0.101)	(0.126)	(0.127)	(0.106)	(0.156)	(427.8)
Book FE	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	No	No	No	No	Yes	Yes
Year-location FE	Yes	Yes	Yes	Yes	No	No
Addtl. Controls	Yes	No	Yes	No	No	No
N	82836	62199	62199	82836	26586	22671

Note: This table reports the robustness of the heterogeneity results to alternate specifications and sample restrictions. Columns (1) through (3) provide zero-inflated Log-OLS estimates that mirror columns (1) through (3) of Table 3. Column (1) adds controls for a book's Google Search volume as well as a dummy variable that equals one if the book has also been digitized on Project Gutenberg before year t. Column (2) drops all books digitized in 2005 or located in Pusey 3, and column (3) adds the demand controls from (1) to the sample from (2). Columns (4) through (6) vary the functional form, mirroring columns (6) through (8) in Table 3. Post-scanned equals one in all years after the book has been digitized. Presales=0 includes books with no sales in 2003 and 2004. Presales>0 describes books between 1 and 500 total sales in 2003 and 2004. And Presales>500 includes all books with more than 500 sales. All models include book and year fixed effects. Columns (1) through (4) additionally interact these year fixed effects with library-location fixed effects. Standard errors are in parentheses, clustered at the book level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table D.4. **Sales Estimates with More Granular Popularity Groups**

	Baseline		Robustness	
	(1) Log-Sales	(2) Log-Sales	(3) Public Domain	(4) Twins
Post-Scanned:				
... × Presales=0	0.0456***	0.0453***	0.0492***	0.0575***
	(0.0123)	(0.0128)	(0.0141)	(0.0165)
... × Presales>0	0.169	0.165	0.180	0.260
	(0.154)	(0.155)	(0.155)	(0.213)
... × Presales>40	0.481***	0.456***	0.460***	0.476***
	(0.132)	(0.133)	(0.133)	(0.182)
... × Presales>490	-0.0175	-0.0296	-0.0217	-0.127
	(0.239)	(0.237)	(0.238)	(0.313)
... × Presales>2020	-0.197	-0.213	-0.205	-0.0607
	(0.157)	(0.155)	(0.155)	(0.250)
... × Presales>7600	-0.0962	-0.131	-0.120	-0.0754
	(0.138)	(0.137)	(0.139)	(0.212)
Book FE	Yes	Yes	Yes	Yes
Year FE	Yes	No	No	No
Year-location FE	No	Yes	Yes	Yes
N	82836	82836	29394	36738

Note: This table provides robustness checks to our popularity cutoff choices. The table mirrors Panel A of Table 4. Presales=0 includes books with no sales in 2003 and 2004. Presales>0 describes books with 1 to 40 total sales in 2003 and 2004. Presales>40 includes all books with 41 to 490 sales in 2003 and 2004. And so on. That is, all popularity groups are mutually exclusive. All columns report results from zero-inflated Log-OLS regressions. Columns (1) and (2) report the baseline regressions. Columns (3) and (4) repeat the sample-related robustness checks from Table 3, using only public domain (digitized) books (3) and including only matched pairs (digitized and not) that are located exactly next to each other (4). Post-scanned equals one in all years after the book has been digitized. All models include book and year-location fixed effects. Standard errors are in parentheses, clustered at the book level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table D.5. **Robustness Checks for Loans Regressions**

	Pre-05 Loans		Sales Sample		Model		
	(1) Log-Loans	(2) Any-Loans	(3) Log-Loans	(4) Any-Loans	(5) Asin(loans)	(6) Poisson	(7) Logit (0/1)
Post-Scanned	-0.0217***	-0.0251***	-0.123***	-0.103***	-0.0656***	-0.484***	-0.501***
	(0.00201)	(0.00202)	(0.00824)	(0.00735)	(0.00196)	(0.0149)	(0.0122)
Book FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	No	No	No	No	No	Yes	Yes
Year-Location FE	Yes	Yes	Yes	Yes	Yes	No	No
N	287523	287523	82836	82836	792054	792054	791028

Note: This table examines the robustness of the loans regressions to sample assumptions. Columns (1) and (2) restrict the sample to books that had at least one loan before the start of digitization (31,947 titles). Columns (3) and (4) estimate the effect on loans using only those books for which we also have sales data (9,204 titles). Log-Loans describes OLS regressions where the dependent variable is the zero-inflated log of loans. Any-Loans describes linear probability models where the dependent variable equals 1 if at least one copy of the book is sold in year t. Post-Scanned equals one in all years after a book has been digitized. Book and year-location fixed effects are included in all models. Standard errors are in parentheses, clustered at the book level. * $p < 0.10$, ** $p < 0.05$, ***$p < 0.01$

Table D.6. **Using Amazon Data to Examine the Effect of Book Digitization**

	Modern Sample			
	(1) Reviews	(2) Reviews	(3) Ln(Rank)	(4) Ln(Rank)
SITB Enabled	269.1***	214.4***	-0.781***	-0.739***
	(57.10)	(65.60)	(0.0415)	(0.0467)
Author FE	Yes	Yes	Yes	Yes
Book FE	—	—	—	—
Year FE	No	Yes	No	Yes
N	11127	10572	10826	10344

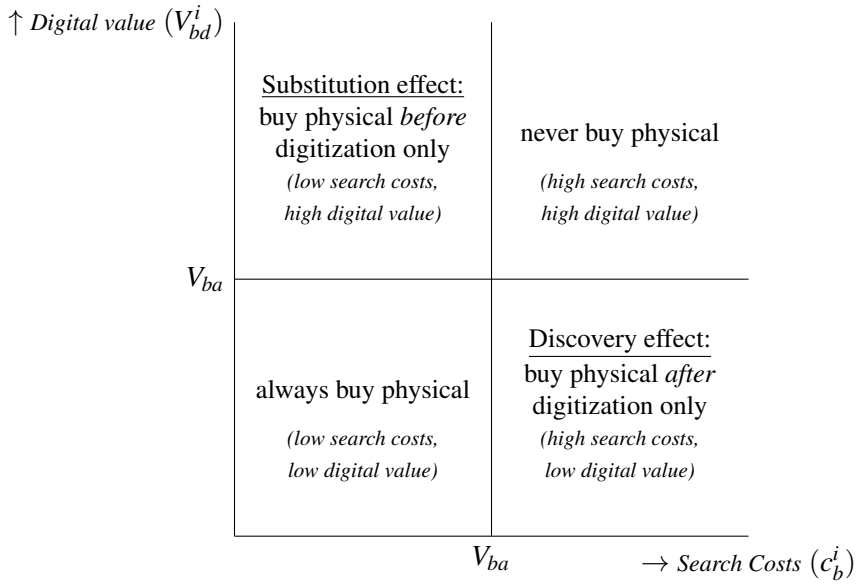
Note: In this table we examine the effect of book digitization using data from Amazon. Cols (1) and (2) present results examining the effect of Google Books digitization on the number of Amazon reviews on a sample of public domain books for which we were able to locate matching Amazon listings. We use the same specification as the baseline regressions, except that the main dependent variable is the number of yearly Amazon reviews. In Cols(3-6), we examine the effect of Amazon's "Search Inside the Book" (SITB) scheme, that scans entire contents of a book and permits readers to search the full-text of the book. We collected a sample of 11,166 in-copyright books by 1775 authors. We examine the effect of SITB in a cross-sectional specification regressing the cumulative number of reviews and ln(Rank) as of Dec 2020 on SITB status, with author fixed effects (Cols 3,5) and/or release year fixed effects (Cols 4,6).

Table D.7. **Impact of Digitization on Prices of New Editions**

	OLS		Log OLS	
	(1) Price	(2) Price	(3) Ln(Price)	(4) Ln(Price)
Post-Scanned	0.181	0.268	-0.00901	-0.00681
	(0.821)	(0.826)	(0.00923)	(0.00929)
Editions		0.0226***		0.000572***
		(0.00478)		(0.0000862)
Book FE	Yes	Yes	Yes	Yes
Year-location FE	Yes	Yes	Yes	Yes
N	275395	275395	275270	275270

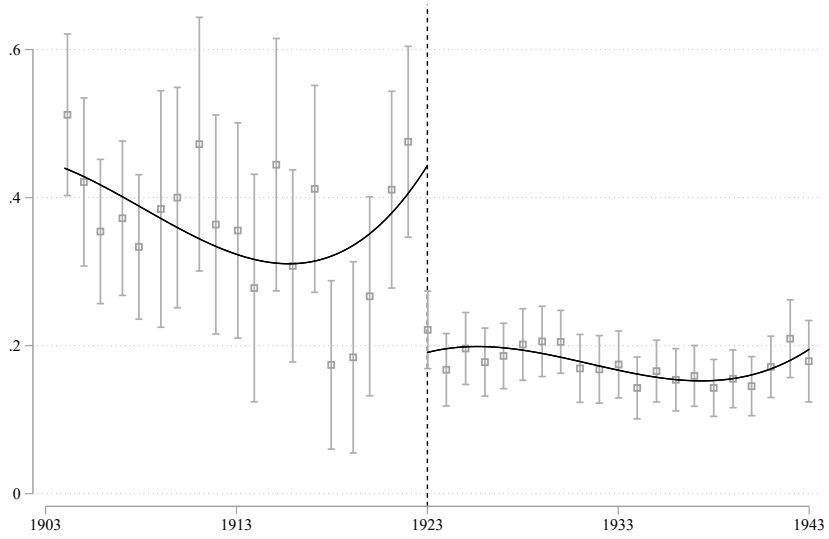
Note: This table presents estimates from OLS and log-OLS models evaluating the overall impact of book digitization on list prices of newly published editions. The estimation is on the edition-level, including all 275,395 editions published between 2003 and 2011 that we matched between the titles in Harvard's library system and the Bowker Books-in-Print directory. Post-Scanned equals one in all years after a book has been digitized. Book and year-location fixed effects are included in all models. Standard errors are in parentheses, clustered at the title level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Figure D.1. **Theoretical Framework: Decision to Consume Physical vs. Digital**



Note: This figure provides an illustration of predictions from the theoretical framework. The framework models an individual customer i's decision to purchase an analog version of the book (a physical copy) as a function of his or her search costs c_b^i (x-axis) and their valuation of the digital copy V_{bd}^i (y-axis) for book b. V_{ba} is the valuation of book b in the physical (analog) format.

Figure D.2. **Annual Estimates of Regression Discontinuity: Likelihood of Increased Sales**



Note: This figure presents event-study estimates of the likelihood that a book sees increased analog sales as a function of its original year of publication. Only those books published before 1923 are digitized due to copyright restrictions. The dependent variable is an indicator that is 1 if analog demand for the book was higher in 2010/11 than in 2003/04, and the independent variables of interest are indicators for the year in which a book was originally published. We plot coefficients for each year of original publication, including 95% confidence intervals (using robust standard errors). A cubic fitted line is included for illustration, and the MSE-optimal bandwidth is chosen.