

Digitization and the Market for Physical Works: Evidence from the Google Books Project[†]

By ABHISHEK NAGARAJ AND IMKE REIMERS*

The free digital distribution of creative works could cannibalize demand for physical versions, but it could also boost physical sales by enabling consumers to discover the original work. We study the impact of the Google Books digitization project on the market for physical books. We find that digitization significantly boosts the demand for physical versions and provide evidence for the discovery channel. Moreover, digitization allows independent publishers to introduce new editions for existing books, further increasing sales. Our results highlight the potential of free digital distribution to strengthen the demand for and supply of physical products. (JEL D12, L82, L86)

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)

Digitization and the advent of the internet have dramatically affected off-line markets for information goods such as books, movies, and music (Brynjolfsson, Hu, and Smith 2003; Forman, Ghose, and Goldfarb 2009; Greenstein, Lerner, and Stern 2013; Waldfogel 2017). The internet shapes physical markets by providing an alternative channel through which content can be consumed, often at very low prices or for free. This phenomenon raises the question of whether and how free digital distribution affects the sales for physical versions of information goods. On one hand, free, digital distribution can lower sales of their physical counterparts by

*Nagaraj: University of California, Berkeley (email: nagaraj@berkeley.edu); Reimers: Cornell University (email: imke.reimers@cornell.edu). C. Kirabo Jackson was coeditor for this article. Saqib Mumtaz Choudhary, Matthew Famiglietti, Scott Schmidt, and Hongyu Yao provided excellent research assistance. We are grateful to attendees of the SERCI Congress, Toronto 2018, Toulouse Digital Economics Conference 2019, UC Berkeley Macro-Lunch, Northeastern University IO-Lunch, NBER Economics of Digitization conference, International Industrial Organization Conference, ZEW Mannheim Research Seminar, MaCCI/ZEW Conference on the Economics of Innovation and Patenting, Munich Summer Institute, Western Economic Association Annual Meeting, Tufts, and Rotterdam School of Management. We further thank Chris Buccafusco, Tristan Botelho, Emily Cook, Daniel Fehder, Michael Kummer, Josh Krieger, Shane Greenstein, Hong Luo, Aruna Ranganathan, Chris Riedl, Pam Samuelson, Mark Seeley, Daniela Sele, Michael Smith, Sameer Srivastava, Scott Stern, Mathijs de Vaan, Joel Waldfogel, and Bruce Weinberg, as well as the referees for useful feedback and comments. We thank Martha Creedon and other members of the staff at the Harvard Libraries for sharing key data used in this paper. All errors are ours.

[†]Go to <https://doi.org/10.1257/pol.20210702> to visit the article page for additional materials and author disclosure statement(s) or to comment in the online discussion forum.

providing consumers a viable substitute. On the other hand, if the digital version facilitates search and discovery of the original product, it can raise physical sales. Given the theoretical ambiguity, we shed light on this question by evaluating the impact of free, digital distribution on the demand for and the supply of physical works using a natural experiment in the context of the Google Books project.

Past work has studied the question of how free distribution affects physical sales largely in the context of file sharing (piracy) for music and movies, finding that the substitution channel likely dominates any potential benefits from discovery. While this work is insightful, it sheds little light on similar questions in the \$25 billion market for books and other printed material. Compared to the markets for music or movies, digitization of books might be more likely to lead to discovery because it can enable full-text search and help consumers discover new content and distributors find additional material to publish. Whether or not online discovery plays a meaningful role in bolstering sales of physical books is an important question for policy. For example, when the Google Books project digitized and freely distributed over 25 million works, it faced legal challenges from publishers and authors who argued that free, digital distribution undermines the market for physical books. More recently, four major publishers are suing the Internet Archive for making books available through a digital lending program using a similar argument.¹ In both cases, proponents of digitization argued that digital distribution can boost sales for physical copies through search and discovery, although their arguments were largely theoretical and backed only by anecdotal evidence. If the discovery hypothesis holds any empirical merit, then such a finding might pave the way for existing and future digitization projects that have attracted considerable legal pushback given concerns around cannibalization. A finding that low-cost digital access to books can boost follow-on innovation and creativity would also be relevant for questions of innovation policy and social welfare (Biasi and Moser 2018; Furman and Stern 2011; McCabe and Snyder 2015).

We begin our analysis by developing a simple theoretical framework that incorporates the substitution and discovery mechanisms when considering the effect of free digital distribution on demand for physical books. The framework clarifies that while the net effect of book digitization is ambiguous, sales could increase if the discovery of a text through a digital channel compensates for the cannibalization of its physical sales. The framework also suggests that the discovery effect should be stronger for less popular books, should apply to nondigitized 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

¹ See https://www.publishersweekly.com/binary-data/ARTICLE_ATTACHMENT/file/000/004/4388-1.pdf.

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). 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. 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. These features make Google Books a prime candidate for identifying a potentially positive effect of digital distribution on physical sales. This setting is also of policy interest because Google Books’ role in cannibalizing sales of existing editions has been debated by scholars and policymakers alike (Samuelson 2009) and has even been presented in front of the US Supreme Court.²

An ideal experiment would randomly provide free, searchable, digital copies of a subset of books and link this variation to changes in physical sales before and after digitization. We come close to this ideal by leveraging a unique natural experiment in which Harvard’s Widener Library (Harvard 2013) 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 trade-off 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 we perform a number of checks to establish the validity of the research design and address any potential concerns.

We combine data from three sources to build a dataset that reports the timing of digitization activity, identifies a comparable set of never-digitized books, and measures off-line demand and supply. First, we obtain data on the shelf-level location of over 500,000 books within the Harvard system between 2003 and 2011, along with information on their loan activity. Second, for a subset of 9,204 books (all books in English with at least four total loans), we obtain weekly US sales data on all related physical editions by manually matching books at Widener with the NPD (formerly Nielsen) BookScan database (Nielsen 2017). Finally, we collect data from the Bowker Books In Print database (Bowker 2017) 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 for and supply of physical editions. Our panel data structure allows for a difference-in-difference design that can incorporate time-library location and book fixed effects.

²The Supreme Court ultimately declined to hear the case.

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 1 sale by 7.7 percentage points. This is our main result. It suggests that at least for the sample of books we study (older and less well known), the Google Books project constituted a net positive for both consumers and publishers. We confirm our findings in a series of robustness checks and tests of the validity of the research design. First, we incorporate time-varying controls at the book level, such as search volume from Google Trends (Google 2018) and availability at Project Gutenberg (Gutenberg 2018). Second, we provide a number of subsample analyses dropping certain books that raise concerns about the exogeneity of their digitization. Third, we create a “twins” sample that consists of pairs of scanned and unscanned books adjacent to each other on the library shelves and hence covering the same subject. Finally, we also collected data on Amazon reviews (Ni 2018) as an alternative measure of physical demand. All results are in line with our baseline findings.

Our framework suggests that for digitization to increase sales of physical works, readers must first discover them online. While we do not possess traffic data on digital readership on Google Books, we provide evidence that digitization through Google Books increased the online use of the digitized books through other channels. In particular, we show that books digitized through Google Books are much more likely to be cited on Wikipedia than their undigitized counterparts. We see this as evidence of the “first stage”: that Google Books provides access to titles that may not otherwise be accessed, and that consumers in fact read (and cite) these books online. Therefore, even though we are largely interested in the diffusion of physical versions, we expect that we are underestimating the effect of digital distribution on diffusion more generally. Further, consistent with our results of the “first-stage” effects and our conceptual framework, we provide evidence that the increase in sales is likely driven by the discovery channel. We find that digitization increases sales significantly for less popular books, and these positive effects disappear for more popular books. Further, digital distribution increases sales for nondigitized 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 works that are likely driven by the discovery of a certain author. In addition, we bring in data from two parallel settings (loans within Harvard and in-copyright books on Amazon) that add further confidence in this finding.

Next, regressing the flow of new book editions on digital availability, we find that digitization increases the number of new editions for books. Although 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. Supporting the discovery mechanism, this effect is largely driven by independent publishers, who have fewer resources for finding good texts than larger publishing houses and university presses. We find that this supply channel is responsible for about 50 percent of the overall physical sales effect, with the remaining half coming from increased demand for existing editions. Our results suggest, first, that digital distribution can stimulate sales through both increased supply of new editions and increased demand for existing

editions, and, second, that free digital distribution can facilitate the entry of smaller publishers, shaping competition in the market for physical information goods.

Our study contributes to two different literatures. First, we speak to past work at the intersection of copyright, digitization, and innovation policy. This work has studied how access restrictions can affect the consumption and diffusion of knowledge (Zhang 2018; Nagaraj 2018; Reimers 2016; Nagaraj, Shears, and de Vaan 2020; Biasi and Moser 2018; Furman, Nagler, and Watzinger 2018). These papers are insightful in that they have shown that loosening access restrictions can help follow-on diffusion. However, this literature has largely ignored the role of spillovers across different channels in driving diffusion. We add to this work by showing how easing access to a digital channel can boost the diffusion of knowledge in both online and off-line settings. Further, in many theoretical models, access restrictions can shape the supply of new works (Landes and Posner 1989), although empirical evidence on this margin is scant. Our result that digital distribution leads to an increase in the number of new editions adds to a small set of papers in this literature that study the impacts of copyright and other restrictions on the supply of new works (Giorcelli and Moser 2020; Reimers 2019).

Second, we also contribute to the literature on the spillovers between online and off-line consumption of media, which has largely focused on the effects of piracy in music and movies (e.g., Rob and Waldfogel 2007; Bai and Waldfogel 2012; Aguiar and Waldfogel 2018). Our results stand in contrast to this literature, which shows that “almost all empirical studies ... find that file sharing has caused a substantial decrease in ... sales” (Smith and Zentner 2016, 435). One exception is Aguiar (2017), who finds that online streaming can stimulate music sales when considering digital purchases. Moreover, by examining cross-channel distribution on the market for books, our work is closely related to Chen, Hu, and Smith (2019), who find no effects of e-book availability (at a positive price and without explicit search functions) on physical sales. By contrast, we show that digital provision can boost rather than cannibalize physical sales when accompanied by search technologies that enable the discovery of new products.

The paper proceeds as follows. We begin by laying out the theoretical arguments for the positive and negative effects of digital provision in Section I. We then describe our data and research design in Section II, followed by a description of the main results, robustness checks, and an exploration of digital use and the discovery mechanism in Section III. We conclude with policy implications in Section IV.

I. Conceptual Framework

Before presenting our empirical analysis, we consider a simple framework to analyze our research question. The framework clarifies that two conditions must be met for digital provision to increase the demand for physical products: (i) the digital product should be an imperfect substitute for the physical product, and (ii) digitization should facilitate consumer discovery of previously unfamiliar content. In addition, the positive effect on sales of physical products can be amplified if publishers also learn about them through digitization and increase supply in the form of new editions. These effects run counter to the forces of cannibalization

that lead consumers to reduce the consumption of physical works and switch to digital alternatives instead. The net effect of digitization, therefore, is ambiguous and depends on the relative magnitudes of these margins, which we call the discovery and substitution channels. We illustrate them in the online Appendix (Figure E.1) and provide a more detailed description here.

A. Demand Effects

The substitution effect is driven by those consumers who would otherwise consume physical copies but switch to digital versions. This is likely to happen when a consumer's search costs are low to begin with and when she has a taste for digital consumption. The dominance of the substitution effect of free digital distribution has found empirical support in the contexts of music (see Danaher, Smith, and Telang 2014 and Oberholzer-Gee and Strumpf 2010 for a review) and movies (Yu et al. 2018; Aguiar and Waldfogel 2018). However, insights from other industries may not apply to the market for books. As compared to music, where digital MP3s may offer a better listening experience than CDs, books might be cumbersome to read online, or only snippets (or partial extracts) of the full text are made available. Both would dampen the substitution effect.

In addition, the discovery effect may also be much more pronounced in the market for books. Digitization can allow for the scanning and searching of the entire text of the book, permitting a much deeper match between content and a customer's preferences. Thus, consumers who were made aware of a book through Google Books' search engine and prefer to purchase physical copies rather than read online may start consuming the physical version for the first time.³ The mass of these consumers who discover a digital version and will purchase a physical version drive the discovery effect. The relative sizes of the two effects determine the net effect of digitization. Given Google Books' full-text search feature and that match quality between content and readers is quite important in the market for books (Ellison and Ellison 2018), the (positive) discovery effect of Google Books may outweigh its (negative) substitution effect.

The trade-off between substitution and discovery further differs for different margins of books and consumers. 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 the nondigitized works of a digitized author as well (Zhang 2018).

³This mechanism is similar to other industries in which (digital) aggregators affect consumption patterns (Kumar, Smith, and Telang 2014; Holtz et al. 2020). For example, news aggregators lead readers to articles they enjoy (Chiou and Tucker 2017; Sismeiro and Mahmood 2018).

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, not the entire text of the book (as is common with "snippet view"), we expect the positive demand effect to remain strong. Our empirical analysis sheds light on these predictions as well.

Note that our discussion so far has not emphasized the role of prices. If consumption of physical versions increases but publishers are forced to reduce prices, then the net effects on revenue might still be negative. However, it is also possible that digital distribution has minimal effects on prices—perhaps because digital distribution attracts a different group of consumers or because publishers do not account for free digital channels when setting prices. Although our theoretical framework does not provide a detailed assessment of the potential channels through which digital distribution could affect prices of physical editions, we will examine this effect empirically.

B. *Supply Effects*

Suppose that publishers publish any content that nets them revenues that are greater than the fixed costs of locating and publishing 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.⁴ 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.⁵ At the same time, free digital provision could increase competition and lower prices, which might reduce publishers' profits per edition and, hence, the likelihood that they will introduce new editions. If the competitive or price effects are minimal, we expect that digital provision will increase the supply of new editions. Further, any positive effects on supply should be especially relevant for small and independent publishers (Nagaraj 2022), who likely face higher costs of locating content as compared to established "major" publishers (Peukert and Reimers 2021).

To summarize, we examine the effects of digital distribution along three margins. First, we examine whether digitization increases or decreases the sales of physical copies, which allows us to evaluate the relative importance of the discovery and substitution effects. As a part of this exercise, we also evaluate the likelihood that digitization increases the digital readership of books. Second, we examine whether any potential positive effects of discovery on off-line sales are stronger for less popular works and might transfer to nondigitized works of digitized authors. Finally, we

⁴ Similarly, Watson (2017) finds that concert performances increase music sales.

⁵ 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.

evaluate whether digital provision allows publishers (especially small and independent ones) to identify new material and introduce new editions.

II. Setting, Data, and Research Design

A. *The Google Books Project: A Brief Background*

The Google Books project (originally known as the Google Print Library Project) was announced by Google in December 2004.⁶ 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 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 United States 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.⁷ 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.”⁸ Google Books' major defense was centered on the idea that browsing books may promote the downstream sales of digitized material.⁹ 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,”¹⁰ and that it was “designed to help you discover books, not read them from start to finish.”¹¹ 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). 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.¹² Note that while Google Books was not the only project digitizing works, it was both the most comprehensive and the most publicized. Two

⁶ See <https://googleblog.blogspot.com/2004/12/all-booked-up.html>.

⁷ See Samuelson (2009) and <https://googleblog.blogspot.com/2008/10/new-chapter-for-google-book-search.html>.

⁸ <http://web.archive.org/web/20190209124325/https://www.authorsguild.org/where-we-stand/authors-guild-v-google/>, originally accessed April 4, 2019.

⁹ See *Authors Guild v. Google* (SDNY 2013), <http://web.archive.org/web/20230203161135/https://h2o.law.harvard.edu/collages/34596>, for more information on the case.

¹⁰ See <http://googlepress.blogspot.com/2004/12/google-checks-out-library-books.html>.

¹¹ <https://web.archive.org/web/20041214092414/http://print.google.com/>.

¹² See <https://www.blog.google/products/search/15-years-google-books/>.

of the largest related projects digitizing public domain works are Project Gutenberg and the Hathi Trust, but they are both smaller and much less popular.

B. *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 their prestigious Widener Library.¹³ 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. Google focused its scanning efforts on multiple different parts of the library at any given point in time. 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.¹⁴ 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 sales, along with book and interacted library shelf location-year fixed effects. Further, the order in which books were scanned was primarily driven by convenience rather than an explicit selection mechanism. We know this through a number of interviews with university officials involved with the Google Books project, including a key 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 Harvard did not "look at it in terms of subject or anything else" (interview with authors, January 15, 2021). He reiterated that since

¹³Other libraries involved in the early phase of digitization, such as the ones 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.

¹⁴If this were the case, we would be less likely to find any effects in our empirical analysis.

a large number of books were involved, there was “absolutely no selection on any base” other than the script of the books (with books in roman script prioritized). This 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 investigate these concerns through our quantitative data analysis, including an analysis of book locations and scan times.

C. Data

The data we obtain from Harvard contain a 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 main dataset from three separate sources.¹⁵ 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 and 2011. Our sample of 88,006 books includes 37,743 (43 percent) that were scanned between 2005 and 2009 and 50,263 (57 percent) that were under copyright and not scanned. Its composition of subject areas is representative of works available at Google Books: about 9 percent of books in US history, 5 percent in economics, and about 3.5 percent 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 percent of total retail sales, although these data do not capture e-book sales.¹⁶ 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 percent of the market; the share was likely lower for older books, such as those in our study. 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 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

¹⁵We introduce supplementary data sources when we use them in later sections.

¹⁶See Berger, Sorensen, and Rasmussen (2010) and <https://www.npd.com/news/category/press-releases/>, accessed June 26, 2018.

of 9,204 titles, or 10.5 percent of the original titles.¹⁷ Of these, 3,267 books (36 percent) were scanned for Google's digitization project.

Third, we collect data on the 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, including their publication dates. We match the 88,006 books in our sample to this database, finding matches for 25,719 unique titles with in-print editions.¹⁸ These data are also helpful to suggest that the lack of e-book sales data is not a problem in our setting, since almost all the books do not have an edition available in e-book form.¹⁹ 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 percent, whereas those that are scanned have a probability of only 8.5 percent before their digitization and 24.1 percent after it. By comparison, books that are never digitized have a probability of being loaned through Harvard's libraries of 17.8 percent, while books that are digitized have a probability of 19.3 percent before their digitization but only 11 percent after their digitization. These differences are indicative of large potential impacts of digitization on demand.

D. Testing the Validity of the Natural Experiment

Our difference-in-difference 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 shelf location \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

¹⁷ Because NPD BookScan does not list books with no recorded sales, we impute zero sales for titles that do not appear in the BookScan database. The results are robust to excluding these titles from the analysis.

¹⁸ 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—for example, dissertations.

¹⁹ Of the books in our sample, only about 3.5 percent of books have at least 1 e-book edition, and 2.1 percent of scanned books released an e-book edition prior to being digitized.

TABLE 1—SUMMARY STATISTICS

	Observations	Mean	Std. dev.	Median	Min	Max
<i>Panel A. Book level</i>						
Scanned (0/1)	88,006	0.43	0.49	0	0	1
Year scanned	37,717	2006.98	1.19	2007	2005	2009
Year of orig. publication	87,808	1910.98	30.61	1925	1560	1943
Total loans (2003–11)	88,006	2.23	5.33	1	1	1,130
Total sales (2003–11)	9,204	4,990.54	56,486.76	0	0	1,965,285
Total editions (2003–11)	88,006	3.21	14.85	0	0	842
Popular (0/1/2)	9,204	0.13	0.46	0	0	2
<i>Panel B. Book-year level</i>						
Postscanned (0/1)	792,054	0.19	0	0	0	1
Loans	792,054	0.25	1	0	0	189
Sales	82,836	554.50	6,839	0	0	626,610
Any-loans (0/1)	792,054	0.17	0	0	0	1
Any-sales (0/1)	82,836	0.16	0	0	0	1
Annual editions	792,054	0.36	3	0	0	542

Notes: 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 1 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 is for books that have been digitized in the time period 2003 to 2011; 37,714 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 = 0 if the title had no sales before the digitization program started (i.e., in 2003 and 2004), = 1 if the title had between 1 and 500 sales in 2003–2004, and = 2 if the title had more than 500 sales before 2005. 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.

random, it is important to examine what 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 1 of 20 possible stacks within Harvard's Widener Library. We are able to match about 81 percent of all scanned books to an exact stack within the library. Using these data, we examine the assertion that the timing of a book's digitization is largely based on its physical location. Figure 1 plots a heat map 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 and 2009 in a given stack. The darker the zone, the higher the percent of books from that stack that were digitized in that time period.²⁰ For example, 73 percent of scanned books in the B West stack were scanned in July 2007, which is indicated by the dark spot on the heat map for this stack. As the series of dark 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 across stacks. Thus, the patterns in Figure 1 support our interpretation that books were digitized based on their stack location.

²⁰The stacks are sorted from bottom to top by the month in which the highest percent of books in their stacks were digitized.

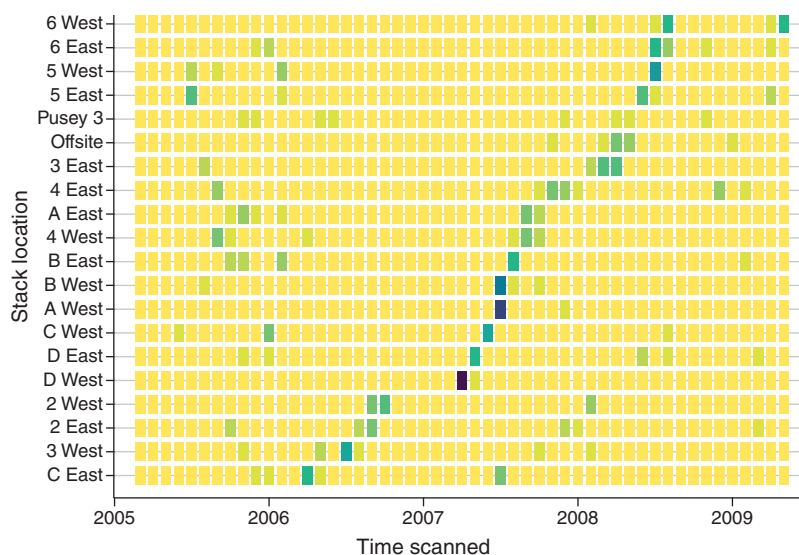


FIGURE 1. TIMING OF BOOK DIGITIZATION BY LIBRARY SHELF LOCATION

Notes: 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 the share of books from that location that 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.

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 book characteristics and predigitization 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 several outcome measures 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, which compare the never-scanned cohort (which has a coefficient of 0 by construction) with cohorts scanned in 2005, 2006, 2007, and 2008–2009, are presented in Figure 2. Panels (i)–(iv) cover book characteristics like the publication year and likelihood of being in the fiction category, as well as the static predigitization demand outcomes: pre-2005 sales and pre-2005 loans. Unsurprisingly, as panel (i) shows, scanned books are published much earlier than never scanned books, but the difference between publication years across the different scanned cohorts is small and seems random. Panels (ii) and (iii) show no significant differences between all cohorts in terms of subject matter and sales.

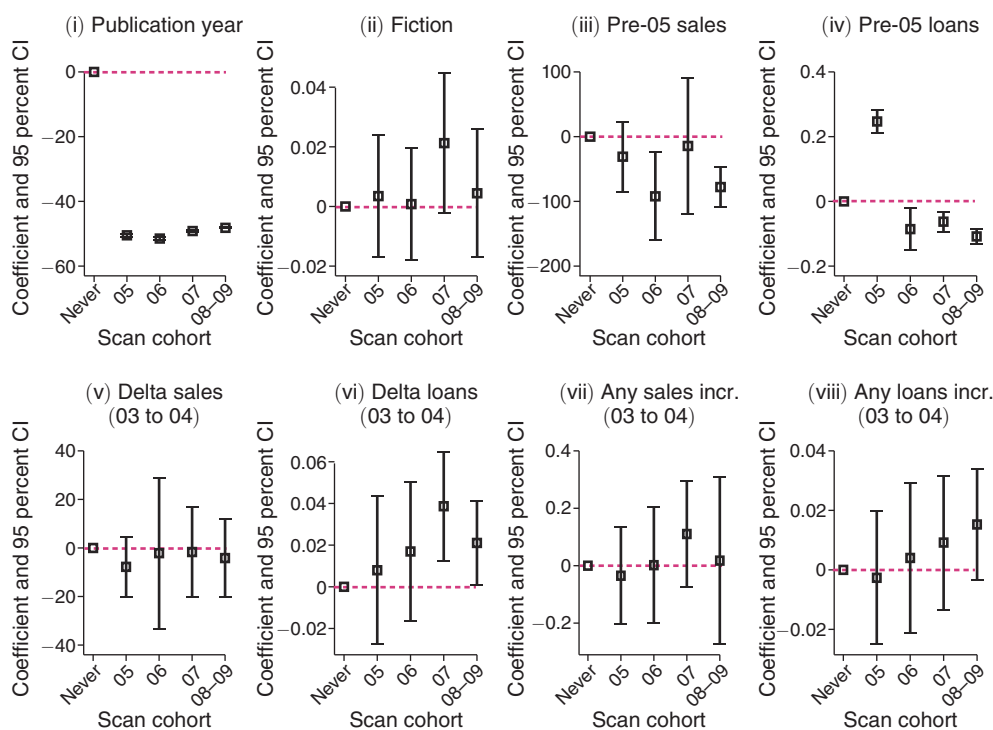


FIGURE 2. COMPARING BOOKS BY SCANNED YEAR

Notes: 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. The dependent variables in the first row are (i) year of publication, (ii) a dummy for whether a book is fiction, (iii) pre-2005 sales, and (iv) pre-2005 loans. Variables in the second row provide a measure of predigitization trends 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 percent confidence intervals (using robust standard errors). The omitted category is books that were not scanned.

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. We find no significant differences across the digitization cohorts.

In sum, our data analyses provide support for our qualitative interviews that suggested that the digitization process was driven by shelf location. However, there could be some concerns for the Pusey 3 stack (which was digitized at different points in time) and for books digitized in 2005 (which have higher levels of pre-2005 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.

III. 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 had been completed.

Figure 3 illustrates how demand changed over the digitization period across different publication cohorts. Panel A plots the share of the books in our sales sample that sold more copies in the 2 years after the digitization period (2010–2011) than in the 2 years before the digitization period (2003–2004), 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 nondigitized 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 2003–2004 to 2010–2011, compared to less than 20 percent of titles that were not digitized. Panel B shows results from a more formal regression approach using event study estimates of the likelihood that a book sees increased sales as a function of the original year of publication (see Section A in the online Appendix). These cross-sectional differences suggest large 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.

A. Main Specification and Results

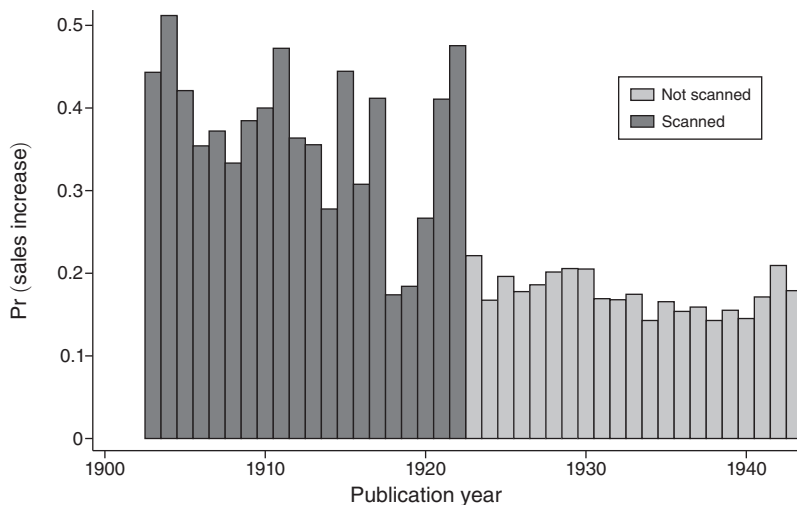
In our main specification, we compare the evolution of sales for titles that were scanned and made available on Google Books with that for titles that were not (yet) digitized in a difference-in-difference setting. Formally, we estimate equations of the form

$$(1) \quad Y_{it} = \alpha + \beta \text{PostScanned}_{it} + \gamma_i + \mu_{it} + \epsilon_{it},$$

where PostScanned_{it} is an indicator that is 1 if book i has been made available on Google Books before year t , γ_i describes book fixed effects, and μ_{it} denotes interacted fixed effects of the year and the book's library location. 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 ($\ln(\text{sales}_{it} + 1)$), where we also examine robustness to adding other constants). In a second set, we use a linear probability model (LPM) where the dependent variable is $\mathbf{1}\{\text{sales}_{it} > 0\}$. That is, we examine the likelihood that a title will have any sales in a given year.

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

Panel A. Mean increase in sales (2003–2011 to 2010–2004)



Panel B. Regression-based estimates

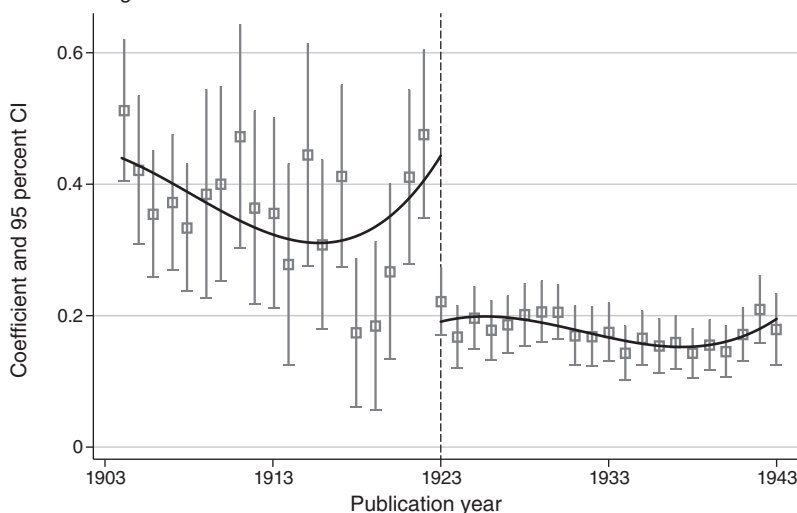


FIGURE 3. COMPARING CHANGES IN SALES FOR PRE-1923 AND POST-1923 BOOKS

Notes: Panel A explores the impact of the digitization program on a cross-sectional sample of English-language books originally published between 1904 and 1942, of which only those published before 1923 were scanned by Google. 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). Panel A plots 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 gray, and those before are indicated in black. Panel B presents event study estimates of the likelihood that physical demand for the book was higher in 2010–2011 than in 2003–2004. 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 percent confidence intervals (using robust standard errors). A cubic fitted line is included for illustration, and the mean square error–optimal bandwidth is chosen.

location fixed effects (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

TABLE 2—BASELINE ESTIMATES FOR THE IMPACT ON SALES

	log-OLS		LPM	
	log-sales (1)	log-sales (2)	Any-sales (3)	Any-sales (4)
<i>Postscanned</i>	0.0480 (0.0125)	0.0466 (0.0130)	0.0782 (0.00481)	0.0770 (0.00487)
Book FE	Yes	Yes	Yes	Yes
Year FE	Yes	No	Yes	No
Year-location FE	No	Yes	No	Yes
Observations	82,836	82,836	82,836	82,836

Notes: 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 $\ln(\text{sales} + 1)$, and columns 3 and 4 report results from LPMs, where the dependent variable is an indicator that is 1 if book i had at least one sale in year t . Postscanned equals 1 in all years after a book has been digitized. 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.

an average of 555 sales per book and year, the estimated increases are also economically significant.²¹ Columns 3 and 4 report similar effects from corresponding LPMs. 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, 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

$$(2) \quad Y_{it} = \alpha + \sum_z \beta_z(\text{scanned})_i \times \mathbf{1}\{z\} + \gamma_i + \mu_{it} + \epsilon_{it}$$

where γ_i and μ_{it} represent book and shelf location \times year fixed effects, respectively; $(\text{scanned})_i$ 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. 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, we set the lag variable to one in the calendar year after the year of digitization.

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 LPM 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, providing support for the validity of our research design. Second, the positive effects on our sales measures seem quite persistent, and they kick in soon after digitization. In online Appendix Figure E.2, we provide versions of these plots that are robust to concerns about heterogeneous

²¹We show in online Appendix Table E.1 that our qualitative results hold when adding smaller constants, although the size of the estimates increases as the constant becomes smaller because adding a smaller constant leads to much larger percentage effects when going from zero sales to positive sales.

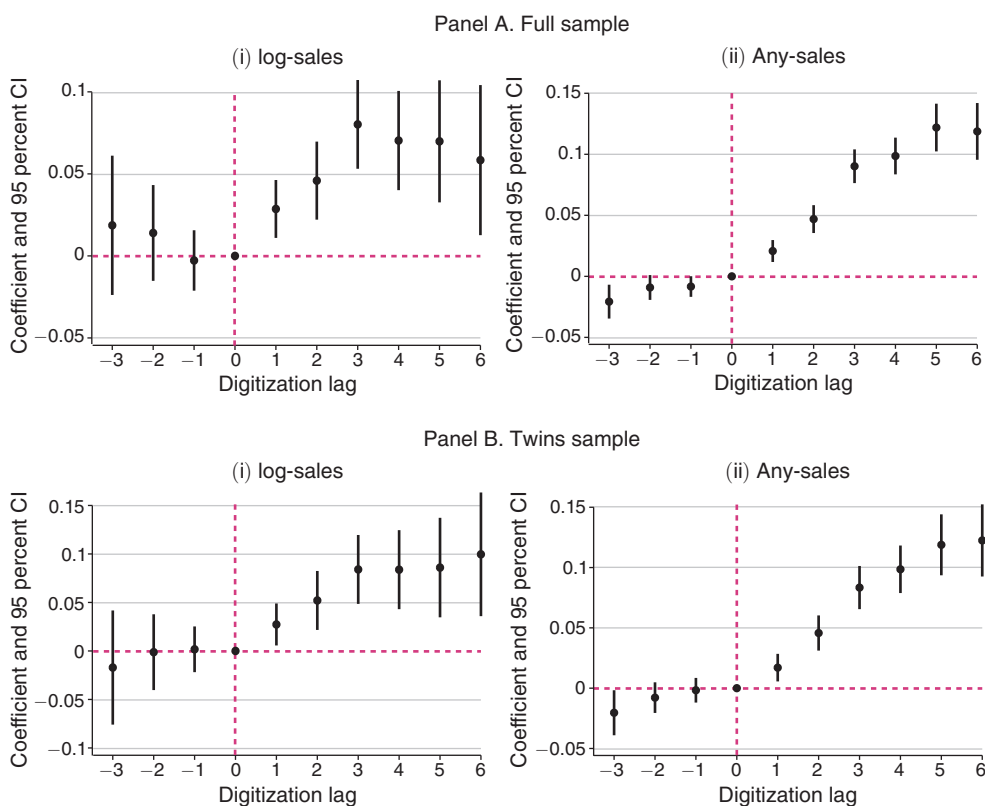


FIGURE 4. TIME-VARYING ESTIMATES OF THE IMPACT OF DIGITIZATION

Notes: This figure provides visual illustrations of the event study specification: $Y_{it} = \alpha + \sum_z \beta_z (\text{scanned})_i \times \mathbf{1}\{z\} + \gamma_i + \mu_t + \epsilon_{it}$, where γ_i and μ_t represent book and year-location fixed effects, respectively, for book i and year t ; $(\text{scanned})_i$ 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 , along with 95 percent confidence intervals.

treatment effects across different treatment cohorts using the estimator developed by Sun and Abraham (2021). In our setting, the results seem quite robust to this concern.

B. 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 two columns examine the potential endogeneity of the timing of digitization. First, we include two additional control variables in our main regressions. We control for potential changes in interest for a title over time by adding each title’s

TABLE 3—ROBUSTNESS CHECKS

	Endogeneity		Sample			Model		
	Controls (1)	Pusey/'05 (2)	Public domain (3)	Twins (4)	Post 1900 (5)	Asin (6)	Poisson (7)	Logit (8)
<i>Postscanned</i>	0.0438 (0.0129)	0.0488 (0.0160)	0.0508 (0.0140)	0.0614 (0.0169)	0.0794 (0.0200)	0.0676 (0.0151)	0.297 (0.153)	0.647 (0.0868)
Book FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	No	No	No	No	No	No	Yes	Yes
Year-loc. FE	Yes	Yes	Yes	Yes	Yes	Yes	No	No
Addl. contr.	Yes	Yes	No	No	No	No	No	No
Observations	82,836	62,199	29,394	36,738	65,457	82,836	26,586	22,671

Notes: This table evaluates the robustness of the baseline regressions to alternate specifications and sample restrictions. Columns 1 through 5 provide zero-inflated log-OLS estimates ($\ln(\text{Sales}_{it} + 1)$). 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 also drops all books digitized in 2005 or located in Pusey 3. Columns 3 through 5 introduce alternate sample restrictions, limiting the sample to books that are in the public domain (3), including only matched pairs (digitized and not) that are located exactly next to each other (4), and using only books that were originally published in or after 1900 (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. *Postscanned* 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.

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.²² 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.²³ 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 is almost unchanged in magnitude, suggesting a digitization-related increase in sales of 4.5 percent.

Second, our analysis of the research design in Section IID 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 located elsewhere. Accordingly, column 2 of Table 3 drops these subsets of titles. Again, the baseline result remains highly significant. In addition, we disambiguate the estimated effects by scan year in online Appendix Table E.2, finding no evidence that our results are driven by any particular digitization cohort.

The next three columns of Table 3 address concerns about the sample in the main specifications. In column 3, we limit the data to books that were in the public

²²The value of 100 for the 2003 normalization is irrelevant for our estimates given the use of year fixed effects.

²³During our period of study, Project Gutenberg had about one-third as many Google searches as Google Books, suggesting its effect is likely smaller.

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 4, 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.²⁴ 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. They also have very similar predigitization 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 4 of Table 3, and the flexible time structure regression from equation (2), in the bottom panel of Figure 4, again with very consistent results. In column 5, we address potential differences across books of different ages. While we show in Table 1 that the oldest book in our dataset was originally published in 1560, only 71 titles were published before 1900. These may have a different longevity that may not be captured by title fixed effects. We therefore drop titles that were originally published before 1900. The estimates from all three specifications are very similar to those from the main sample, if not stronger.

In our main analyses, we also make certain assumptions about the functional forms and error terms. We explore the robustness of our results to these assumptions in the remaining columns of Table 3. The inverse hyperbolic sine specification, $\text{asinh}(\text{sales}_{it})$, in column 6, offers an alternative way of addressing 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. In online Appendix Figure E.3, we further examine whether digitization affects the probability that other sales thresholds—from 1 to 1,000 annual sales—are surpassed. The probability of selling more units increases for small thresholds, up to five units.

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 on the Amazon platform from Ni, Li, and McAuley (2019). We use these reviews to proxy for a book's demand at Amazon, and we estimate the effect of digitization on these reviews.²⁵ Our results suggest a digitization-related increase of 3.97 annual Amazon reviews, relative to an average of 1.07 reviews per year for books in our dataset.

²⁴For example, this approach drops all unscanned books that are located between two other unscanned books.

²⁵We find matches for 559 scanned books in our sales sample. For these books, we construct a balanced sample between the years 1997 and 2018, for a total of 12,298 observations. We then estimate a version of equation (1) with annual reviews as the dependent variable.

C. Digitization and Digital Use

Why might digital distribution boost off-line sales? Our conceptual framework suggests that digital distribution can increase discovery through increased digital readership, which then boosts off-line sales. We investigate this proposition in two steps. In this subsection, we consider the first stage: we examine whether digitized books see meaningful readership in digital channels and whether access restrictions curtail digital readership. In Section IIID we evaluate the discovery channel more directly.

Ideal data on digital use would come from Google Books directly. But such information is not available at the book level. Instead, we examine digital use through another channel: use on Wikipedia. As documented in prior work (Nagaraj 2018), digitized material on Google Books is often referenced in Wikipedia. Wikipedia editors looking for authoritative sources of information might discover it on Google Books, and then incorporate relevant material on Wikipedia and, helpfully for us, also include a citation to the particular title at the bottom of the page. Adding a citation for a title on Google Books is a high bar; it means that the title was being read (at least by a set of interested Wikipedia editors) and that it was being cited as a source for an article that deserves encyclopedic inclusion. Therefore, if we do find that titles on Google Books are cited often on Wikipedia and that access restrictions lower citations, this would provide strong evidence that digital distribution greatly boosts digital readership. For example, the Wikipedia page on George Washington makes extensive references to the Google Books page for the *Journals of the Continental Congress* (Ford et al. 1904, and digitized by Google Books), containing important biographical details about Washington's life.²⁶ These citations are strong evidence to suggest that the digitization and full-text access to Ford et al. (1904) led to not just digital readership but also follow-on reuse via a citation on Wikipedia.

We rely on data from Singh, West, and Colavizza (2021), who extracted over 29.3 million citations from 6.1 million Wikipedia articles as of May 2020. Using this dataset, we identify all instances of books in our sample being cited on Wikipedia via a fuzzy matching procedure on book titles. When considering a very high threshold of what we consider to be a match, we find that 4,295 books (or about 5 percent of our sample) have at least 1 citation on Wikipedia. Note that we expect the rate of digital readership to be many times higher than the citation rate. This provides reassuring evidence that the Google Books digitization indeed leads to digital use, a necessary first step for our proposed discovery mechanism.

Beyond establishing digital use, we compare whether titles in our sample are less likely to be cited on Wikipedia when digital access is restricted. In Figure 5 we compare the probability of Wikipedia citations to the 56,855 books in our sample published between 1904 and 1942 (20 years before and after 1923). Note that we count citations to book titles, so if a nondigitized book is being read and cited, even if it does not have a Google Books link, we should be able to detect a citation in our

²⁶ See https://en.wikipedia.org/wiki/George_Washington and <https://books.google.com/books?id=zMSAAAAYAAJ>.

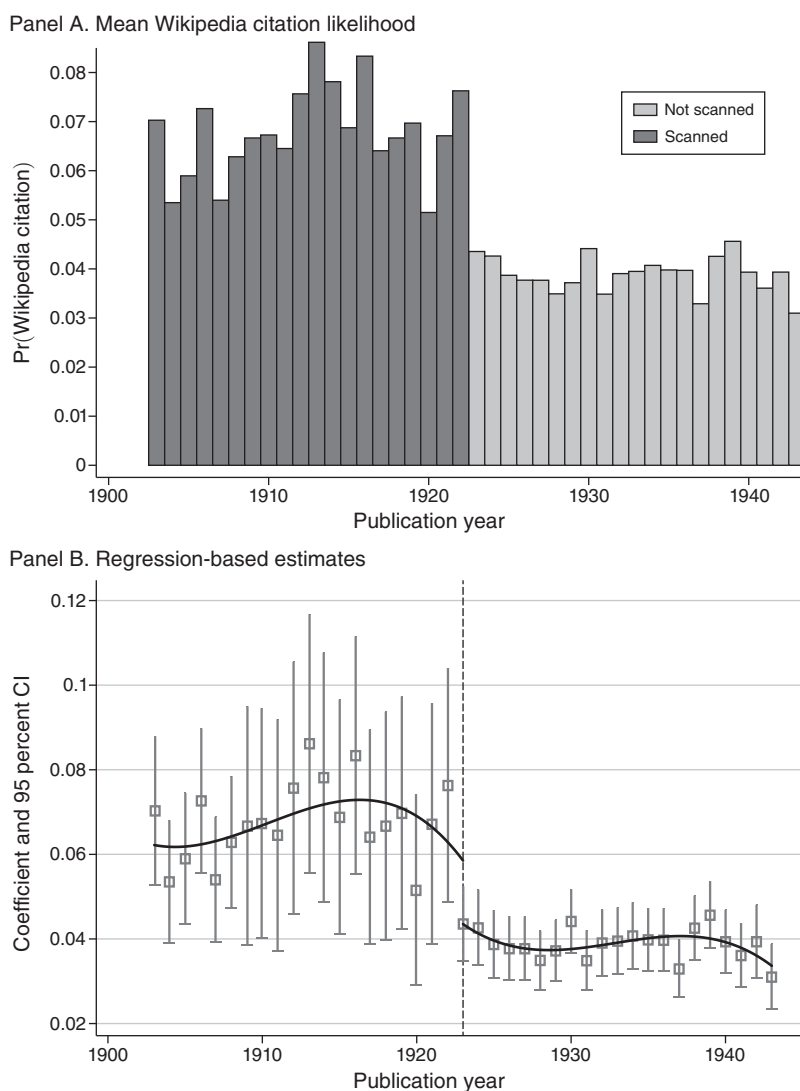


FIGURE 5. LIKELIHOOD OF WIKIPEDIA CITATION FOR PRE-1923 AND POST-1923 BOOKS

Notes: This figure explores whether digitized books get used more than nondigitized books. We include all books in our original sample of 88,006 books that were published between 1904 and 1942, for a sample of 56,855 books for this analysis. Books published before 1923 are digitized. For each book, we measure whether or not it is cited at least once on Wikipedia, either via a direct link to the book on Google Books or via a fuzzy title match. We then plot the share of books in each publication year that have at least one cite on the y-axis and the publication year on the x-axis. Books published after 1923 are indicated in gray, and those before are indicated in black.

data. Even after accounting for citations in this broad way, Figure 5 shows that there is a sharp drop in citations around the 1923 cutoff, after which titles are much less likely to be digitized and made available on Google Books. This pattern provides compelling evidence to suggest that improving digital access substantially boosts digital use.

Section D in the online Appendix presents a variety of additional analyses that corroborate these findings. First, rather than split the books by publication year as in Figure 5, we compare citations by direct information on scan status for our sample of 88,006 books and find similar patterns. Then, rather than restrict to the sample of digitized books from Harvard, we consider 39,439 citations to over 28,554 Google Books URLs for titles published between 1904 and 1943 and document that a much greater proportion of citations go to books with full-text access and those published before 1923. Finally, rather than rely on Wikipedia for reuse information, we collect data on backlinks across the internet—information on whether a Google Books page was linked from a blog, newspaper article, online forum, etc.—for a broader accounting of digital use beyond Wikipedia for the books in our baseline sample. We find that 16.3 percent of pre-1923 books have at least one backlink, while this number drops to 8.8 percent for books published post-1923.

Combined, our data on Wikipedia citations to Google Books as well the additional analyses presented in the online Appendix help to establish the first stage, that digital distribution fuels digital readership.

D. Digitization and Discovery

The sections above suggest that digitization—when coupled with improved searchability—can in fact improve digital readership and increase the sales of physical copies. Our theoretical framework links these increases to facilitated discovery and decreases in search costs. We now investigate 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 whether 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.

Heterogeneous Effects by Book Popularity.—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 2003–2004): books with no sales (91.9 percent of the sales sample), books with 1 to 500 sales (3.1 percent), and books with more than 500 sales (5 percent). We then repeat our baseline estimations, interacting our postscanned variable with indicators for each popularity group. The results are reported in panel A of Table 4. The first two columns mimic the first two columns in Table 2, and the remaining columns mimic the first two sample-based robustness checks from Table 3. We find a small but statistically significant positive effect in the first group (presales = 0): the estimated coefficient of 0.0456 can be interpreted roughly as a 4.6 percentage point increase in the likelihood of having any sales. We observe large and significant positive effects—sales increases of over 30 percent—of digitization on books that had previously been relatively obscure (presales between 1 and 500), and no statistically significant effect on the most popular

TABLE 4—EXPLORING THE DISCOVERY MECHANISM

	Baseline		Robustness	
	log-sales (1)	log-sales (2)	Public domain (3)	Twins (4)
<i>Panel A. Effects by the book's popularity</i>				
<i>Postscanned:</i>				
... × <i>Presales</i> = 0	0.0456 (0.0123)	0.0453 (0.0128)	0.0492 (0.0141)	0.0575 (0.0165)
... × <i>Presales</i> > 0	0.324 (0.103)	0.309 (0.104)	0.319 (0.103)	0.364 (0.142)
... × <i>Presales</i> > 500	−0.112 (0.101)	−0.133 (0.100)	−0.125 (0.100)	−0.0860 (0.149)
Book FE	Yes	Yes	Yes	Yes
Year FE	Yes	No	No	No
Year-location FE	No	Yes	Yes	Yes
Observations	82,836	82,836	29,394	36,738
	In copyright		All books	
	log-sales (1)	Any-sales (2)	log-sales (3)	Any-sales (4)
<i>Panel B. Spillovers to the author's other books</i>				
<i>Postscanned ×</i>				
... <i>this book</i>			0.0553 (0.0316)	0.0878 (0.00959)
... <i>other book</i>	0.103 (0.0460)	0.0346 (0.0108)	0.0663 (0.0354)	0.0299 (0.00891)
Book FE	Yes	Yes	Yes	Yes
Year-location FE	Yes	Yes	Yes	Yes
Observations	19,224	19,224	29,799	29,799

Notes: 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 first two 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 panel B 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 LPMs. 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.

works.²⁷ Assuming that the discovery effect is less pronounced among books that are already well known, these results provide some of the first evidence that discovery plays a role in determining the effect of digitization.

²⁷We report results from additional robustness checks analogous to Table 3 in Table E.3 the online Appendix. We also explore more granular popularity cutoffs in Table E.4 and Figure E.4 in the online Appendix.

Spillovers across Authors.—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’s 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 reports 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 effects of the free digital provision through Google Books on other books, quite comparable in size to the direct effects.

Discovery versus Substitution.—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 channels. 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 regressions based on equation (1), using library loans as the dependent variable. Showing results from the zero-inflated log-OLS model as well as an LPM, the first two columns imply that digitization decreases loans on average by about 5 percent and decreases the probability that a book is checked out at Harvard at all in a year by 6.1 percentage points.²⁸ 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. Nevertheless, 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 partial access to in-copyright books that were digitized, Harvard did not participate

²⁸We report results from robustness checks in the online Appendix (Table E.5).

TABLE 5—EFFECTS ON HARVARD LIBRARY LOANS

	Main effect		Consumer groups	
	log-OLS	LPM	Non-Harvard	Harvard
<i>Postscanned</i>	−0.0511 (0.00152)	−0.0613 (0.00170)	−0.0362 (0.00121)	−0.0157 (0.000964)
Book FE	Yes	Yes	Yes	Yes
Year-location FE	Yes	Yes	Yes	Yes
Observations	792,054	792,054	792,054	792,054

Notes: 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 an LPM (2). The next two columns disambiguate the effects from the log-OLS model for loans from non-Harvard affiliated consumers (3) and Harvard affiliated consumers (4). All models include book and year-location fixed effects. Standard errors are in parentheses, clustered at the book level.

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 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.²⁹ We find significant differences in our demand measures: books that are included in 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 73 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. The fact that authors and publishers willingly utilize the program for their more successful books is also telling.

E. 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 identify, create, and publish more and higher-quality copies of public domain books. We examine the effect of digitization on the supply of 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.

²⁹We provide more detail on the program and how it helps us identify the discovery function of digitization in Section B of the online Appendix.

TABLE 6—THE ROLE OF NEW EDITIONS

	Effect on editions				Effect on sales	
	New eds. (1)	Majors (2)	Indies (3)	Cumulative (4)	log-sales (5)	log-sales (6)
<i>Postscanned</i>	2.091 (0.130)	0.0302 (0.00581)	2.061 (0.127)	5.061 (0.328)		0.0259 (0.0132)
<i>Cumulative editions</i>					0.00920 (0.00166)	0.00408 (0.000635)
Book FE	Yes	Yes	Yes	Yes	Yes	Yes
Year-location FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	82,836	82,836	82,836	82,836	82,836	82,836

Notes: 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 nonmajor 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. Postscanned 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.

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). Moreover, 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, which further suggests that the positive effects on sales translate into meaningful increases in publisher revenues.³⁰ Naturally, the increase in the number of *new* editions implies 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 the cumulative number of available editions with the book's digitization status. That is, column 4 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 editions

³⁰See Section C in the online Appendix.

on downstream demand. We find that the positive effect of digitization on log-sales remains significant but decreases 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.

IV. Discussion

Digital distribution and search offer a powerful tool to reshape how content is produced, discovered, consumed, and distributed. In this paper, we show that free digital distribution may increase rather than decrease the sales of physical works. In our setting, the positive effect of free digital provision on demand is stronger for more obscure books and spills over to a digitized author's nondigitized books. Our results are consistent with the idea that digitization allows readers to discover new works and consume them in physical form. 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.

Our results 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 and 1936 (Reimers 2019). Not only could these books be made available for digital access, but digitization might also increase the availability and sales of their physical editions. Therefore, our results help strengthen the value proposition of mass-digitization projects such as Google Books, Project Gutenberg, the Internet Archive, or the Hathi Trust that have faced legal pushback. Further, historical textual material can often be of significant scholarly and research interest. Access to past knowledge has been shown to greatly improve innovation, creativity, and entrepreneurship in other domains (Biasi and Moser 2018; Furman, Nagler, and Watzinger 2018; McCabe and Snyder 2015). Projects like the Google Books project that provide access to past works in digitized form could allow future innovators and creators to discover new bases of knowledge that might unlock large benefits in terms of follow-on innovation and creativity.

While the general welfare implications of the Google Books Project are beyond the scope of this study, we can provide some back-of-the-envelope calculations to assess the costs and benefits of the program. Some estimates suggest that it cost Google about \$400 million to scan 25 million works (Somers 2017). This suggests an average cost of about \$16 per book, although this number probably varies depending on the difficulty of scanning a work. Our estimates suggest that digitization increased the sales of the average book in our sample by about 5 percent. Since the mean book sold about 5,000 times in the time period of our study, this would translate to about 250 more sales per book. Our price data suggest that the average book sells for about \$30 and that this price did not decrease after a title's digitization. Past work suggests that publishers have a profit margin of about 50 percent (Reimers and Waldfogel 2017). These estimates suggest that publishers would make about $\$250 \times 15 = \$3,750$ per book, against a cost of about \$16 for digitizing it. These

numbers suggest a very large benefit of Google Books digitization for publishers vis-à-vis the cost to Google. Note that our analysis suggested broader benefits in terms of digital readership and citations for digitized books, which are unpriced and therefore harder to factor into traditional cost-benefit analyses (Brynjolfsson, Collis, and Eggers 2019).

Our findings also inform how changes in the representation and aggregation of works shape firm policies 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 distribute digital versions of their works to boost physical sales. While we find that even full-text access can stimulate demand, the positive impacts of “snippet” access could be even stronger, because it allows for the full power of text-based discovery but mutes incentives for cannibalization. As such, our results are aligned with an internal study conducted by Amazon that found a 9 percent increase in sales for books that had the SITB feature enabled.³¹ Even if publishers are wary of digital distribution in general, an optimal policy may consist of 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 work is not without limitations. First, our intention is not to say that search-enabled digital distribution will necessarily increase physical sales. Rather, we 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 off-line 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, deepening our understanding of how consumers and publishers discover content in a world with both physical and digital channels remains an exciting topic for future research.

³¹<https://press.aboutamazon.com/2003/10/amazon-com-announces-sales-impact-from-new-search-inside-the-book-feature>.

REFERENCES

- Aguiar, Luis.** 2017. "Let the Music Play? Free Streaming and Its Effects on Digital Music Consumption." *Information Economics and Policy* 41: 1–14.
- Aguiar, Luis, and Joel Waldfogel.** 2018. "As Streaming Reaches Flood Stage, Does It Stimulate or Depress Music Sales?" *International Journal of Industrial Organization* 57: 278–307.
- Bai, Jie, and Joel Waldfogel.** 2012. "Movie Piracy and Sales Displacement in Two Samples of Chinese Consumers." *Information Economics and Policy* 24 (3–4): 187–96.
- Berger, Jonah, Alan T. Sorensen, and Scott J. Rasmussen.** 2010. "Positive Effects of Negative Publicity: When Negative Reviews Increase Sales." *Marketing Science* 29 (5): 815–27.
- Biasi, Barbara, and Petra Moser.** 2018. "Effects of Copyrights on Science—Evidence from the U.S. Book Republication Program." NBER Working Paper 24255.
- Bowker.** 2017. "Books in Print." <https://www.bowker.com/en/products-services/books-in-print> (accessed August 3, 2023).
- Brynjolfsson, Erik, Avinash Collis, and Felix Eggers.** 2019. "Using Massive Online Choice Experiments to Measure Changes in Well-Being." *Proceedings of the National Academy of Sciences* 116 (15): 7250–55.
- Brynjolfsson, Erik, Yu (Jeffrey) Hu, and Michael D. Smith.** 2003. "Consumer Surplus in the Digital Economy: Estimating the Value of Increased Product Variety at Online Booksellers." *Management Science* 49 (11): 1580–96.
- Chen, Hailiang, Yu Jeffrey Hu, and Michael D. Smith.** 2019. "The Impact of E-book Distribution on Print Sales: Analysis of a Natural Experiment." *Management Science* 65 (1): 19–31.
- Chiou, Lesley, and Catherine Tucker.** 2017. "Content Aggregation by Platforms: The Case of the News Media." *Journal of Economics and Management Strategy* 26 (4): 782–805.
- Danaher, Brett, Michael D. Smith, and Rahul Telang.** 2014. "Piracy and Copyright Enforcement Mechanisms." *Innovation Policy and the Economy* 14: 25–61.
- Ellison, Glenn, and Sara Fisher Ellison.** 2018. "Match Quality, Search, and the Internet Market for Used Books." NBER Working Paper 24197.
- Ford, Worthington Chauncey, Gaillard Hunt, John Clement Fitzpatrick, Roscoe R. Hill, Kenneth E. Harris, Steven D. Tilley, and Library of Congress Manuscript Division.** 1904. *Journals of the Continental Congress, 1774–1789*. Washington, DC: U.S. Government Printing Office.
- Forman, Chris, Anindya Ghose, and Avi 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.
- Furman, Jeffrey L., and Scott Stern.** 2011. "Climbing atop the Shoulders of Giants: The impact of Institutions on Cumulative Research." *American Economic Review* 101 (5): 1933–63.
- Furman, Jeffrey L., Markus Nagler, and Martin Watzinger.** 2018. "Disclosure and Subsequent Innovation: Evidence from the Patent Depository Library Program." NBER Working Paper 24660.
- Giorcelli, Michela, and Petra Moser.** 2020. "Copyrights and Creativity: Evidence from Italian Opera in the Napoleonic Age." *Journal of Political Economy* 128 (11): 4163–210.
- Google.** 2007. "World's Oldest Publisher Stays at the Cutting Edge with Google Book Search." <https://books.google.com/intl/en-US/googlebooks/cup.html> (accessed August 3, 2023).
- Google.** 2018. "Google Trends." <https://www.google.com/trends> (accessed August 3, 2023).
- Greenstein, Shane, Josh Lerner, and Scott Stern.** 2013. "Digitization, Innovation, and Copyright: What Is the Agenda?" *Strategic Organization* 11 (1): 110–21.
- Gutenberg.** 2018. "Project Gutenberg." <https://www.gutenberg.org> (accessed August 3, 2023).
- Harvard.** 2013. "Harvard Library Data Extract." <https://library.harvard.edu> (accessed August 3, 2023).
- Holtz, David, Benjamin Carterette, Praveen Chandar, Zahra Nazari, Henriette Cramer, and Sinan Aral.** 2020. "The Engagement-Diversity Connection: Evidence from a Field Experiment on Spotify." *EC '20: Proceedings of the 21st ACM Conference on Economics and Computation*: 75–76.
- Kumar, Anuj, Michael D. Smith, and Rahul Telang.** 2014. "Information Discovery and the Long Tail of Motion Picture Content." *MIS Quarterly* 38 (4): 1057–78.
- Landes, William M., and Richard A. Posner.** 1989. "An Economic Analysis of Copyright Law." *Journal of Legal Studies* 18 (2): 325–63.
- McCabe, Mark J., and Christopher M. Snyder.** 2015. "Does Online Availability Increase Citations? Theory and Evidence from a Panel of Economics and Business Journals." *Review of Economics and Statistics* 97 (1): 144–65.
- Nagaraj, Abhishek.** 2018. "Does Copyright Affect Reuse? Evidence from the Google Books Digitization Project." *Management Science* 64 (7): 2973–3468.

- Nagaraj, Abhishek.** 2022. "The Private Impact of Public Data: Landsat Satellite Maps and Gold Exploration." *Management Science* 68 (1): 564–82.
- Nagaraj, Abhishek, Esther Shears, and Mathijs de Vaan.** 2020. "Improving Data Access Democratizes and Diversifies Science." *Proceedings of the National Academy of Sciences* 117 (38): 23490–98.
- Nagaraj, Abhishek, and Imke Reimers.** 2023. "Replication Data for: Digitization and the Market for Physical Works: Evidence from the Google Books Project." American Economic Association [publisher], Inter-university Consortium for Political and Social Research [distributor]. <https://doi.org/10.3886/E175701V1>.
- Nielsen.** 2017. "NPD BookScan." <https://bookscan.npd.com> (accessed December 1, 2022).
- Ni, Jianmo.** 2018. "Amazon Review Data." <https://nijianmo.github.io/amazon/index.html> (accessed February 20, 2021).
- Ni, Jianmo, Jiacheng Li, and Julian McAuley.** 2019. "Justifying Recommendations Using Distantly-Labeled Reviews and Fine-Grained Aspects." *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*: 188–97.
- Oberholzer-Gee, Felix, and Koleman Strumpf.** 2010. "File Sharing and Copyright." In *Innovation Policy and the Economy*, Vol. 15, edited by William R. Kerr, Josh Learner, and Scott Stern, 19–55. Chicago: University of Chicago Press.
- Peukert, Christian, and Imke Reimers.** 2021. "Digitization, Prediction and Market Efficiency: Evidence from Book Publishing Deals." Unpublished.
- Reimers, Imke.** 2016. "Can Private Copyright Protection Be Effective? Evidence from Book Publishing." *Journal of Law and Economics* 59 (2): 411–40.
- Reimers, Imke.** 2019. "Copyright and Generic Entry in Book Publishing." *American Economic Journal: Microeconomics* 11 (3): 257–84.
- Reimers, Imke, and Joel Waldfogel.** 2017. "Throwing the Books at Them: Amazon's Puzzling Long Run Pricing Strategy." *Southern Economic Journal* 83 (4): 869–85.
- Rob, Rafeaul, and Joel Waldfogel.** 2007. "Piracy on the Silver Screen." *Journal of Industrial Economics* 55 (3): 379–95.
- Samuelson, Pamela.** 2009. "Legally Speaking: The Dead Souls of the Google Book Search Settlement." *Communications of the ACM* 52 (7): 28.
- Singh, Harshdeep, Robert West, and Giovanni Colavizza.** 2021. "Wikipedia Citations: A Comprehensive Data Set of Citations with Identifiers Extracted from English Wikipedia." *Quantitative Science Studies* 2 (1): 1–19.
- Sismeyro, Catarina, and Ammara Mahmood.** 2018. "Competitive vs. Complementary Effects in Online Social Networks and News Consumption: A Natural Experiment." *Management Science* 64 (11): 5014–37.
- Smith, Michael D., and Alejandro Zentner.** 2016. "Internet Effects on Retail Markets." In *Handbook on the Economics of Retailing and Distribution*, edited by Emek Basker. Northampton, MA: Edward Elgar Publishing.
- Somers, James.** 2017. "Torching the Modern-Day Library of Alexandria." *The Atlantic*, April 20.
- Sun, Liyang, and Sarah Abraham.** 2021. "Estimating Dynamic Treatment Effects in Event Studies with Heterogeneous Treatment Effects." *Journal of Econometrics* 225 (2): 175–99.
- Waldfogel, Joel.** 2017. "How Digitization Has Created a Golden Age of Music, Movies, Books, and Television." *Journal of Economic Perspectives* 31 (3): 195–214.
- Waldfogel, Joel, and Imke Reimers.** 2015. "Storming the Gatekeepers: Digital Disintermediation in the Market for Books." *Information Economics and Policy* 31: 47–58.
- Watson, Jeremy.** 2017. "What is the Value of Re-use? Complementarities in Popular Music." Unpublished.
- Yu, Yanan, Hailiang Chen, Chih-Hung Peng, and Patrick Y. K. Chau.** 2018. "The Causal Effect of Subscription Video Streaming on DVD Sales: Evidence from a Natural Experiment." Unpublished.
- Zhang, Laurina.** 2018. "Intellectual Property Strategy and the Long Tail: Evidence from the Recorded Music Industry." *Management Science* 64 (1): 24–42.