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

Abhishek Nagaraj
UC Berkeley-Haas
nagaraj@berkeley.edu

Imke Reimers
Northeastern University
i.reimers@northeastern.edu

April 15, 2019

Abstract

The age of digitization promised to deliver a centralized, digital repository of all knowledge. Copyright holders, however, concerned about reduced demand for physical works, have blocked the realization of this vision. We investigate the effect of digitization on demand for physical works using novel data tracking the timing of the digitization of individual books from Harvard University's libraries through the Google Books project. Digitization hurt loans within Harvard but increased sales of physical editions by about 35%, especially for less popular works. Rather than cannibalizing demand, digitization might benefit copyright holders through increased discovery of less popular works.

*The authors thank Saqib Mumtaz Choudhary, Matthew Famiglietti, and Scott Schmidt for excellent research assistance. Chris Buccafusco, Tristan Botelho, Daniel Fehder, Michael Kummer, Josh Krieger, Shane Greenstein, Hong Luo, Aruna Ranganathan, Chris Riedl, Pam Samuelson, Mark Seeley, Sameer Srivastava, Mathijs de Vaan, Joel Waldfogel and attendees of the SERCI Congress, Toronto 2018, Toulouse Digital Economics Conference 2019, UC Berkeley Macro-Lunch, Northeastern University IO-Lunch and NBER Economics of Digitization conference provided useful 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.

1 Introduction

Digitization and the advent of the Internet have dramatically transformed the creation and distribution of information goods such as books, movies, and music (Greenstein et al., 2013; Waldfogel, 2017). Not only has digitization facilitated the creation of new products, but it has also significantly expanded access to the back-catalog of physical works. Much like a modern-day Library of Alexandria, there is the real possibility that the Internet could serve as a repository of all pre-existing knowledge in digital form (Samuelson, 2011). Further, in the case of books, digitization allows readers to search through the full-text of printed material, dramatically lowering search costs and facilitating discovery. Increased access to past knowledge through digitization could boost follow-on innovation, productivity, and creativity (Furman and Stern, 2011; Furman et al., 2018; Biasi and Moser, 2018; Williams, 2013). This idea is not just a pipe dream. Efforts led by for-profit organizations such as the Google Books project, as well as non-profit groups like the Hathi Trust and the Internet Archive, have spent tens of millions of dollars digitizing the world's books and making them easier to search. By last count, over 25 million books had already been digitized through these efforts (Somers, 2017).

Despite the technological progress and financial investment, there still exists no single digital repository where the sum of human knowledge can be accessed freely and at low cost.[1] This result can be largely attributed to legal considerations, especially copyright challenges from traditional publishers and authors that have been fiercely litigated and even presented in front of the US Supreme Court.[2] Copyright holders are concerned about the possibility that digitized versions would serve as substitutes for material in print, thus hurting an industry that made over $40 billion in revenue in 2008.[3] Some copyright holders are aware of the potential value of digital distribution channels but are opposed to unlicensed use of their materials under "fair use" since they are not compensated directly. In contrast, proponents of digitization argue that, among other benefits, many works have become obscure over time, and easily accessible digital versions can increase awareness and discovery, thereby increasing demand.[4] If the negative effects of digitization on the demand for printed works are indeed low, digitization might be a win-win for consumers, publishers, and authors, and could pave the way for large-scale and unlicensed digitization of works without costly

[1] Google, for example, "all but shut down its scanning operation" (Somers, 2017), and even though Google Books is operational, it does not provide unfettered access to most books ever published.

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

[3] See Michael Healy, Book Industry Study Group, Books and e-Books: Some Industry Numbers, at the D is for Digitize Conference at the NY Law School 2009, http://www.nyls.edu/innovation-center-for-law-and-technology/iilp-archive/iilp-conferences/d_is_for_digitize/.

[4] In line with both arguments, a 2012 survey of users of a Norwegian digitization effort found that 20% of respondents purchased a book after first finding it on the depository, whereas 18% reported that they did not (Josevold, 2016).

negotiations with individual copyright holders. In this paper, we move beyond the theoretical debates and attempt to provide empirical estimates of whether and to what extent digitization harms demand for physical works.

We focus on the Google Books project, which was launched in 2004 with a vision to digitize all works ever created and which was the leading contender to become the modern-day Library of Alexandria. This project digitized existing works and made them available online for free in "snippet" form for in-copyright material and in full-text form for public domain books (our focus here). The site also included an interface for in-text search to facilitate discovery. The project was launched in partnership with Harvard University's Widener library (as well as a handful of others), which provided public domain books and texts to be digitized by Google. In this context, the net effect of Google Books on demand depends on the number of pre-existing consumers who switch to digital consumption (lowering demand) and the number of new consumers, who discover works online and then purchase a physical edition (increasing demand).

Our research design builds on proprietary data from Harvard which includes the date on which a particular work was digitized by Google. Book digitization at Widener Library took over five years, and while the order of digitization of books was not deliberately randomized, it was not explicitly selective either and did not prioritize the most interesting books for early digitization. Rather, our fieldwork suggests that digitization proceeded on a "shelf-by-shelf" basis and was driven by convenience. We exploit the timing of variation between digitized books, as well as between books that were or were not scanned, to evaluate the impact of digitization on the demand for physical works in a difference-in-differences analysis with book and year fixed effects. In a parallel set of analyses, we exploit a sharp discontinuity in digitization across books published before and after 1923 (with the latter set not digitized) for additional estimations.

We combine data from three sources for our analysis. First, we collect data on library loans within the Harvard system between 2003 and 2011. Even though our data includes over half-a-million works, most of these are never loaned, and so our analyses focus on 88,006 books that had at least one loan in the sample period.[5] Second, for a subset of 9,204 books (in English with at least four loans), we obtain weekly US sales data on all related editions from the NPD (formerly Nielsen) BookScan database.[6] Finally, we also collect data on publications of new editions from the Bowker Books-In-Print database to examine the effect of digitization on bringing out-of-print books back to life.

Our results suggest that the impact of digitization on demand is negative when considering internal read-

[5] Our results are not affected if we focus on only those books with at least one loan before 2005, but that sample is smaller.

[6] The sales data must be manually collected and matched by hand, which restricts the size of this sample.

ership at Harvard but is positive when considering sales. In our preferred specification, digitization lowers the number of checkouts within Harvard by about 37% but increases sales by about 34%, or about 200 units annually at the mean. Alternatively, digitization reduces the probability that a book will be loaned at least once in a year by about 33% but increases the likelihood of at least one sale by 92%. We further find that the positive effect of digitization on sales is significantly greater for less popular books. These findings suggest the discovery effects of digitization are likely to be significant for less popular books and for readers without access to alternate ways to search for books (such as Harvard's internal catalog system). Further, we find that digitization leads to an uptick of new in-print editions through other publishers, likely making these books more easily available to consumers. However, this improved availability (and any associated change in book prices) explains only a small fraction of the increased overall sales due to digitization, further reinforcing the importance of the discovery mechanism.

Our results provide much-needed empirical evidence in ongoing legal debates around the viability of mass digitization of works for copyright holders who argued for a "market replacement" effect of Google Books in *Author's Guild vs. Google (2013)*. Our finding that digitization, on the whole, increased sales suggests that concerns about cannibalization of physical works are likely overblown. In contrast, copyright holders might be able to increase discovery and demand through digitization, especially for less popular and out-of-print works, even when they might not be compensated directly by third-party repositories relying on fair use.

Beyond the specific legal debates around mass digitization, we also contribute to the emerging literature on the economic effects of copyright law. This work has shown that stronger copyright law can incentivize creativity (Giorcelli and Moser, 2016) and increase prices (Li et al., 2018). However, stronger copyrights can also harm the ability of follow-on creators to build on pre-existing work (Heald, 2007; Reimers, 2019) in contexts as diverse as Wikipedia (Nagaraj, 2017), music (Watson, 2017), and academic research (Biasi and Moser, 2018). Our work adds to this literature by focusing on the policy debate around the mass digitization of books and examining the impacts of weakening copyright restrictions through digitization. Our work also adds to the literature examining the effect of the internet on the market for books (Ellison and Ellison, 2018; Brynjolfsson and Smith, 2000), and the individual strategies of publishers in a digital setting (Chen et al., 2018; Reimers, 2016). Finally, while a parallel literature in music has debated the effects of piracy (Smith and Telang, 2012), our paper is related to work that shows that less-stringent piracy restrictions might be particularly helpful for less popular publishers (Kretschmer and Peukert, 2014; Zhang, 2016).

2 The Google Books Project: A Brief Background

The Google Books project (originally known as the Google Print Library Project)[7] was announced by Google in December 2004. At the project's inception, Google partnered with Harvard University's library (along with a few other key partners), to digitally scan books from their collections. Soon after these works were scanned, they were made available on the Google Books website for the general public.[8] The site provided access to the full-text of public domain books (including books published in the USA 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.[9] Authors and publishers were particularly concerned about Google's unauthorized distribution of their works, a problem that was compounded by 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."[10] Google Books' major defense was centered on the idea of fair use and the notion that browsing books may promote the downstream sales of digitized material.[11] 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."[12] and that it was "designed to help you discover books, not read them from start to finish."[13] In other words, since the market for books is one where match quality is quite important (Ellison and Ellison, 2018), potential readers might discover a book relevant to them via Google Books, and then go ahead and purchase a physical copy of the book, thereby stimulating sales. Empirical evidence on either of these two claims was sparse.

The suits were eventually settled (publishers) or rejected (authors), but the process lasted over a decade before an appeal by the Authors Guild was rejected in the Second Circuit. As an upshot of these intense legal

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

[8] While Google Books was not the only project digitizing works, it was by far the most comprehensive and ambitious project of its time.

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

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

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

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

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

battles, as of 2018, the Google Books project remains a far cry from the original ambitions behind its founding. As one commentator puts it, "Somewhere at Google there is a database containing 25-million books and nobody is allowed to read them" (Somers, 2017). Similar projects launched by non-profit organizations, such as the Hathi Trust, have also been hampered by legal restrictions.

In the legal debates, the question of whether the digitization of books can reduce demand for physical works depends on two counteracting forces: the discovery effect of Google Books due to increased awareness and searchability, and the substitution effect of digital distribution as a competitor for existing, physical products. We clarify this theoretical tension with a simple conceptual framework in Appendix A, as depicted visually in Appendix Figure D.1. As we clarify in this framework, some consumers who consumed analog copies will now switch to digital versions (perhaps because they have a taste for digital consumption, and because it is free), driving the substitution effect. On the other hand, some consumers may start consuming the analog version for the first time, after digitization lowers search costs for books. This is likely to happen if they were made aware of a book through Google Books' search engine and prefer to purchase physical copies rather than read online. This second mass of consumers will drive the discovery effect. The net effect of digitization depends on the magnitude of these two margins.

We argue that the tradeoff between substitution and discovery differs for different margins of books and consumers. Notably, for popular books, already well-known to consumers (e.g. *The Wealth of Nations*), the substitution effect is likely to dominate. On the other hand, obscure books are likely to benefit from discovery, but not face the costs of substitution. The effect of Google Books on demand should therefore be more positive for less popular books. Similarly, 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, we expect digitization to have a greater positive effect on demand when considering market-wide sales, while for loans within Harvard, the effect is likely much smaller, and even negative. Our empirical analysis sheds light on both of these predictions as well.

3 Data and Research Design

3.1 Google Books and Harvard Libraries' Natural Experiment

Given the unclear legal environment around digitization and copyright when the project began, Harvard's participation in the Google Books project was limited to works from Harvard's prestigious Widener Library

whose copyright was deemed to have expired. 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 and were therefore provided for scanning. Since this cutoff date would not change until much after the digitization was completed, books from after 1923 were not digitized.

The digitization of Harvard's public domain books proceeded as follows. The Google Books project set up a scanning facility in the Greater Boston area to process the books from the Harvard libraries. For the purposes of the scanning effort, Google Books was assigned a special library patron code, and books were "loaned" to Google under this special code to be taken to the scanning facility. Once the book was scanned, it was returned to the library and made available on the Google Books website after a short delay, usually within a few weeks (personal communication, December 2011).

Our natural experiment relies on the fact that the scale of Google's scanning project at Harvard implied that the total duration of the project was over five years (from 2005 to 2009), after which it was shut down. In our baseline analysis, we rely on this variation in the timing of the scanning project to estimate the impact of digitization on eventual readership and sales, along with book and year fixed effects. Further, our conversations with Harvard librarians indicate that the order in which books were scanned was driven by convenience rather than an explicit selection mechanism. Specifically, books were scanned on a shelf-by-shelf and wing-by-wing basis until all out-of-copyright books in the relevant sections were processed, a claim we examine through analyzing differences in pre-trends between digitized and non-digitized works.

3.2 Data

The data we obtain from Harvard contain the entire record of over 250,000 books from the Harvard libraries' holdings that were scanned, as well as a similar number of works published between 1923 and 1943 that were not scanned. For these books, we possess proprietary checkout data, which allows us to infer total loans within Harvard, as well as the date when the book was checked out by Google Books for digitization. Using these data, we construct our baseline sample, which includes all 88,006 books that were checked out at least once between 2003-2011.

Second, we obtain access to NPD (formerly Nielsen) BookScan, which provides weekly sales information for printed books. NPD tracks book sales using scanner data from a large panel of retail booksellers including major bookstore chains, discount retailers such as Costco, and major online retailers like Amazon. They

claim to track about 85% of total retail sales.[14] 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 editions for each title. Given the tedious data collection process, we search for sales data for the subset of all English-language books in the underlying dataset with at least four loans, for a total of 9,204 titles. Because NPD BookScan does not explicitly list books with no recorded sales, we impute zero sales for titles that do not explicitly appear in the BookScan database. The results are robust to excluding these titles from the analysis.

Third, we collect data on the number of in-print editions of all works from the Bowker Books-in-Print database. This database tracks all registered editions of a particular work that are available in print. We match the 88,006 books in our sample to this database, finding matches for almost 25,000 unique titles with in-print editions. 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 the digitization on the demand for physical works within Harvard (loans) and in the market (sales). 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. These data contain loans information for 88,006 books, including 50,263 books that were deemed not in the public domain and hence not digitized, and 37,743 that were. Of the ones that were digitized, 5,764 were scanned in 2005, 7,449 in 2006, 8,769 in 2007, 13,207 in 2008, and 2,546 in 2009. In any given year, an average book has 0.25 loans, sells about 554 copies, and adds 0.36 editions, although the median value for all three outcomes is zero. Over the entire sample, books are loaned on average 2.23 times and have sales of almost 5000. These data are summarized in Table 1 Panels A (book-level) and B (book-year level).

4 Results

We measure loans and sales for titles that were scanned and made available on Google Books, and we compare the evolution of these measures with that of titles that were not (yet) digitized in a difference-in-differences setting. Formally, we estimate equations of the form

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

[14] See Berger et al. (2010) and https://tinyurl.com/y94qpsqt, accessed June 26, 2018.

where $PostScanned_{it}$ is an indicator that is 1 if book i has been made available on Google Books before year t, and γ_i and μ_t are book and year fixed effects, respectively. The dependent variable, Y_{it}, denotes book- and year-specific measures of demand (loans and sales). To account for the discrete nature and low average values of the dependent variables, we assume that the error term ε_{it} follows a Poisson distribution, and we therefore estimate the model in a maximum likelihood estimation. Poisson models for count data rely on quite weak assumptions and are appropriate (and quite commonly used) in our context with skewed dependent variables (Wooldridge, 1999). In a second set of baseline analyses, we estimate a similar specification using a linear probability model (LPM) where Y_{it} is either $1(sales_{it} > 0)$ or $1(loans_{it} > 0)$. That is, we examine the likelihood that a book will have any sales or loans in a given year after digitization. In robustness checks, we estimate similar models using OLS and Log-OLS specifications as well.

4.1 Loans and Sales

Table 2 Panel A displays the results from the Poisson (columns 1 and 2) and LPM (columns 3 and 4) specifications. Columns 1 and 3 show that digitization through Google Books significantly decreases demand at Harvard's libraries. Column 1 suggests that making the book available on Google Books decreases Harvard library loans by about 36.7% ($= e^{-0.457} - 1$). The likelihood that a book will have any loans decreases by 6.3 percentage points. At the mean across scanned books before their digitization (19.3%), this corresponds to a 32.7% increase in the likelihood of a loan each year. These results are in line with the prediction that for consumers with access to search technology (such as Harvard students and faculty), digitization largely displaces demand for physical alternatives.

Now we turn to examining the sales results, which estimate the *market-wide* effects of digitization based on the subsample of books for which we have sales information. Unlike loans, sales through traditional channels increase after digitization. The Poisson estimates (column 2) suggest an increase in sales of 34.5% ($= e^{0.297} - 1$) per year due to digitization (p-value 0.053). Similarly, the likelihood of making at least one sale increases by 7.8 percentage points (p-value less than 0.001), or by 91.8% at the mean. Overall, this analysis indicates that, while digitization might reduce demand for those with access to Harvard's libraries, the *market-wide* impact is positive and significant.

4.2 Timing of the Impact

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

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

where γ_i and μ_t represent book and time fixed effects, respectively, $(scanned)_i$ 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.[15] We estimate this equation corresponding to the linear probability models estimated in Table 2.

Figure 1 illustrates the estimates for β_z for the loans and sales outcomes. Two points are clear from this analysis. First, there are no significant pre-trends in terms of the likelihood of being loaned or sold in a given year between books that already were and are yet to be scanned. If anything, loans for yet-to-be scanned books are increasing just prior to digitization. Second, it seems like the negative effect of digitization on loans and the positive effect on sales are quite persistent and long-lasting, and kick in soon after digitization. These graphical results lend support to the interpretation that digitization had a causal impact on reduced demand within Harvard and increased sales through other channels.

4.3 Separating Effects for Popular Books

The positive market-wide impact of digitization suggests a strong role for the discovery effect in driving demand. We investigate the discovery mechanism further by examining whether the Google Books project differently impacts books of different popularity levels. We repeat the analyses from Table 2, with an additional interaction term of the Post-Scanned variable with an indicator that equals 1 for books in the 90$^{\text{th}}$ percentile of pre-2005 loans. This definition identifies all books that were checked out at Harvard's libraries more than once in 2003-2004 as popular. Here, we implicitly assume that the interest within Harvard is a proxy for the book's popularity, an assumption supported by the fact that within the sales sample, popular books sell twice as much as less popular books on average. Our results are largely robust to different definitions of popularity, as shown in Appendix B.

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

Table 2 Panel B presents estimates from this exercise. Consistent with expectations, the effect of digitization on increasing demand is smaller for popular books. Columns 1 and 3 indicate that all digitized books see a significant decrease in loans compared to books that were not digitized or digitized later. However, the impact is significantly more negative for popular works, which experience a decrease of about 62% ($= e^{-0.362-0.599} - 1$) compared to a decrease of 30% for less popular books, according to the point estimates in column 1. Alternatively, while the likelihood that a book is checked out reduces by about 3.4 percentage points for less popular books, for more popular books this estimate is over 28 percent points.

When considering sales, less popular books experience an almost 42% increase in sales – consistent with facilitated discovery outweighing substitution – whereas the impact on sales of more popular books is muted, with an (imprecisely) estimated increase in sales of about 16%. In other words, for the vast majority of books, the positive effect on demand through discovery outweighs the negative effect of substitution, and we find no support for authors' and publishers' concerns about the cannibalization of their work. When considering the likelihood of making at least one sale, both popular and less popular books benefit from digitization. In fact, it seems like a larger share of popular books benefits from digitization in terms of making at least one sale. However, this result is likely due to the fact that our estimates measure percentage point changes, rather than percentages. We view this result as less instructive for policy given that making at least one sale is less meaningful for publishers of popular books, which sell on average over 700 copies a year, according to our data.

4.4 Robustness Checks

To further bolster our baseline estimation, we present results from two sets of robustness checks, including estimating alternate specifications and accounting for the role of new editions in driving increased demand.

Alternate Specifications: Table 3, Panel A (Cols 1-4) presents estimates from OLS specifications where the dependent variable is either Y_{it} (Col 1-2), or $Ln(Y_{it}+1)$ (Col 3-4). These results are largely in line with the baseline analysis, although some of the estimates (especially in the OLS models) are imprecisely estimated given the skewed nature of the data. The remaining two columns (5-6) present results from our baseline Poisson specifications in which we drop extreme outliers – eight titles with more than 900,000 lifetimes sales.[16] The results again remain consistent with the baseline analysis, although the magnitude on the sales coefficient is smaller.

[16]This cutoff represents a rather large gap in lifetime sales between included and excluded books. Results are robust to other cutoffs.

Accounting for Editions: Next, we consider the possibility that Google Books enables publishers to create and publish more and higher-quality copies of public domain books, in turn allowing consumers to buy editions that did not exist previously. We first estimate the impact of digitization through Google Books on the availability of *new* editions in the market, analogous to the estimations in the baseline analyses. We then examine whether the changes in sales can be attributed to improved availability through new editions.

Table 3 Panel B presents these results. Columns 1-2 show the impact of digitization on a book's availability and find that titles become available in more new editions after their digitization, with an average increase of 71% ($= e^{.538} - 1$) (Col. 1). When considering the stock of editions, titles become more likely to be available at all, with an increase in that likelihood of 18.6 percentage points, as against a baseline of 11.3% (Col 2).

Table 3 Panel B (Cols 3-6) estimates the impacts of digitization on use, now controlling for the number of newly introduced editions of the work.[17] Columns 3-4 control for new editions linearly, while columns 5-6 provide more flexible controls, including dummy variables for each number of new editions, combining observations with more than eight editions in one group.[18] These estimates show that new editions do have a small direct impact on demand. However, despite controlling for this effect, the impact of digitization on loans and sales remains statistically significant and of similar magnitude.

While we do not possess systematic price data for our entire sample, we do obtain prices at the edition level from the Bowker data. In Appendix B we explore whether an increase in sales we found could be offset by a decrease in prices. We document a small decrease in prices which is an order of magnitude lower than our estimate for the increase in sales. Further, this decrease in prices is largely related to the increase in the number of editions, and therefore controlling for editions also helps to adjust for any possible price channel driving the sales effect.

Additional Robustness Checks: Beyond the above concerns, our difference-in-differences analysis is based on additional choices regarding the sample of books (all books checked out at Harvard at least once between 2003 and 2011) and the definition of popular works (a work is popular if it was checked out more than once between 2003 and 2004). We show in Appendix B that our results are robust to these choices as well. Further, in unreported analyses, we find that the estimates are robust to controlling for differential time trends, i.e. grouping books into the library sections (e.g. "Literature") and including group by year fixed effects.

[17] We treat all titles without a match in the Bowker database as a title with zero editions. However, our results are unchanged if these titles are dropped.

[18] This aggregation affects 0.7% of all data.

4.5 Exploiting the Discontinuity Around 1923

While our analysis so far relies on the timing variation in the digitization of books across Harvard, we can exploit another source of variation to evaluate the effects of digitization on demand in this context. Specifically, books published in the United States before 1923 are in the public domain, while those published later might have copyright protections. Accordingly, when working with Google, Harvard allowed US books to be digitized only if they had been published before 1923. We exploit this sharp cutoff to examine whether loans and sales of books published right before this cutoff changed considerably compared to books published right after, once the digitization process had been completed.

Here, we consider 39,949 US books that were published 20 years before and after 1923. For each of these books, we calculate the total number of loans and sales before digitization (i.e., for the years 2003 and 2004) and the equivalent figure in the years 2010 and 2011, after the digitization project concluded. We then calculate an indicator for whether the loans and sales for each of these books increased across the two periods. The average value of this indicator by publication year is presented in Figure 2 for loans (i) and sales (ii). The lighter bars represent the likelihood of an increase in loans/sales for digitized books, and the darker bars represent the likelihood of an increase for books that were not digitized.

Two points are worth noting about this figure. First, there seems to be a discontinuous and sharp change in the increase in loans and sales around 1923, with loans increasing and sales decreasing when moving from the (digitized) pre-1923 group to the (non-digitized) post-1923 group. Further, the darker and the lighter bars are relatively flat within their particular groups, suggesting that the impact of digitization is conditional on whether a book was scanned or not, and not on other unobserved factors that might drive outcomes (such as publication year) and confound the causal interpretation of our estimates.

While we prefer our baseline panel estimates given the relatively clean pre-trends and the reliance on the entire sample, the descriptive analysis in Figure 2 is quite reassuring. In Appendix C, we estimate regression-discontinuity models inspired by this descriptive analysis. The results from this analysis are qualitatively identical and quantitatively very similar to those from our main specifications.

5 Discussion

Our empirical analysis suggests that digitization, through discovery, may increase rather than decrease demand for physical works. This finding has important implications for ongoing legal and policy debates on the design of copyright law for the digital age. First, our evidence contradicts the popular notion that digitization necessarily harms demand for physical works. Therefore, our results help strengthen the value proposition of mass-digitization projects such as Google Books, the Hathi Trust or the Internet Archive. While previous negotiations have tried to weigh the benefits to society against the harm to copyright holders, we find that this tradeoff might be relevant only when there is little potential for additional discovery through digitization, for example for very popular books. Our findings also point to the utility of digitization for individual, less popular authors looking to boost their readership.

While our evidence comes from the digitization of public domain books (published before 1923), it also speaks to debates about the digitization of newer, in-copyright works. Our evidence comes from providing the full text of public domain books in digital form, whereas for in-copyright works the debate is about providing "snippets" of relevant text. Given that we find no meaningful substitution effect even when the entire book is provided in digital form, and given that the creation of new editions – likely the result of full texts being made available – does not drive the positive results, the overall positive effects we estimate could be even stronger for in-copyright works where only 20% of the text is provided.

While we advance the broader debate on the impact of digitization in the market for books, it is important to acknowledge the limitations of our study. First, we focus on the digitization of a sample of public domain books from just one (albeit prominent) library. It is possible that these effects could be different for a more general sample of contemporary books. Second, results from our study might not generalize to a world where e-book consumption and complementary devices (such as the Kindle) are changing tastes for analog consumption. Finally, the overall welfare effect depends on how digitization changes the dynamic incentives of authors and publishers to produce and finance new work. Our estimates do not measure the elasticity of this important margin.

In sum, our study clarifies the important role of digitization in enabling discovery and helping copyright holders increase the sales of physical editions of their works. We hope these findings help rekindle the debate about digitizing printed material and getting humanity one step closer to a digital library of Alexandria.

References

Berger, J., A. T. Sorensen, and S. J. Rasmussen (2010, March). Positive Effects of Negative Publicity: When Negative Reviews Increase Sales. *Marketing Science 29*(5), 815–827.

Biasi, B. and P. Moser (2018, December). Effects of Copyrights on Science. SSRN Scholarly Paper ID 2542879, Social Science Research Network, Rochester, NY.

Brynjolfsson, E. and M. D. Smith (2000). Frictionless commerce? a comparison of internet and conventional retailers. *Management science 46*(4), 563–585.

Chen, H., Y. J. Hu, and M. D. Smith (2018). The Impact of E-book Distribution on Print Sales: Analysis of a Natural Experiment. *Management Science*.

Ellison, G. and S. F. Ellison (2018). Match quality, search, and the internet market for used books. Technical report, National Bureau of Economic Research.

Furman, J. L., M. Nagler, and M. Watzinger (2018). Disclosure and Subsequent Innovation: Evidence from the Patent Depository Library Program. Technical report, National Bureau of Economic Research.

Furman, J. L. and S. Stern (2011, August). Climbing atop the Shoulders of Giants: The Impact of Institutions on Cumulative Research. *American Economic Review 101*(5), 1933–1963.

Giorcelli, M. and P. Moser (2016, December). Copyrights and Creativity: Evidence from Italian Operas. SSRN Scholarly Paper ID 2505776, Social Science Research Network, Rochester, NY.

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

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

Josevold, R. (2016). A National Library for the 21st centuryknowledge and cultural heritage online. *Alexandria 26*(1), 5–14.

Kretschmer, T. and C. Peukert (2014). Video killed the radio star? online music videos and digital music sales.

Li, X., M. MacGarvie, and P. Moser (2018). Dead Poets' Property-How Does Copyright Influence Price? *The RAND Journal of Economics 49*(1), 181–205.

Nagaraj, A. (2017, July). Does Copyright Affect Reuse? Evidence from Google Books and Wikipedia. *Management Science 64*(7), 3091–3107.

Reimers, I. (2016). Can private copyright protection be effective? evidence from book publishing. *The journal of law and economics 59*(2), 411–440.

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

Samuelson, P. (2009). Legally Speaking: The Dead Souls of the Google Book Search Settlement. *Communications of the ACM 52*, 28.

Samuelson, P. (2011). The Google Book Settlement as Copyright Reform. *Wis. L. Rev.*, 479.

Smith, M. D. and R. Telang (2012). Assessing the academic literature regarding the impact of media piracy on sales.

Somers, J. (2017). Torching the Modern-Day Library of Alexandria. *https://www.theatlantic.com/technology/archive/2017/04/the-tragedy-of-google-books/523320/*.

Waldfogel, J. (2017). How Digitization Has Created a Golden Age of Music, Movies, Books, and Television. *Journal of Economic Perspectives 31*(3), 195–214.

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

Williams, H. L. (2013). Intellectual property rights and innovation: Evidence from the human genome. *Journal of Political Economy 121*(1), 1–27.

Wooldridge, J. M. (1999). Distribution-free estimation of some nonlinear panel data models. *Journal of Econometrics 90*(1), 77–97.

Zhang, L. (2016). Intellectual property strategy and the long tail: Evidence from the recorded music industry. *Management Science 64*(1), 24–42.

6 Tables and Figures

Tables

Table 1. **Summary Statistics**

Panel A: Book-Level

	Mean	Std. Dev.	Median	Min	Max
Scanned (0/1)	0.43	0.49	0.00	0	1
Year Scanned	2006.98	1.19	2007.00	2005	2009
Total Loans (2003-11)	2.23	5.33	1.00	1	1130
Total Sales (2003-11)	4990.54	56486.76	0.00	0	1965285
Total Editions (2003-11)	3.21	14.85	0.00	0	842
Popular (0/1)	0.12	0.32	0.00	0	1

Panel B: Book-Year Level

	Mean	Std. Dev.	Median	Min	Max
Post-Scanned (0/1)	0.19	0.39	0.00	0	1
Loans	0.25	0.89	0.00	0	189
Sales	554.50	6839.49	0.00	0	626610
Any Loans (0/1)	0.17	0.37	0.00	0	1
Any Sales (0/1)	0.16	0.37	0.00	0	1
Annual Editions	0.36	2.90	0.00	0	542

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

Table 2. **Estimates for the Impact on Loans and Sales**

Panel A. Overall Impact

	Poisson		OLS	
	(1) Loans	(2) Sales	(3) Any-Loans	(4) Any-Sales
Post-Scanned	-0.457	0.297	-0.0629	0.0782
	(0.0175)	(0.153)	(0.00167)	(0.00481)
Book FE	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes
N	792054	82836	792054	82836

Panel B. Heterogenous Effects by Popularity

	Poisson		OLS	
	(1) Loans	(2) Sales	(3) Any-Loans	(4) Any-Sales
Post-Scanned	-0.362	0.349	-0.0341	0.0670
	(0.0206)	(0.190)	(0.00170)	(0.00625)
Post Scanned x Popular	-0.599	-0.201	-0.246	0.0244
	(0.150)	(0.221)	(0.00368)	(0.00900)
Book FE	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes
N	792054	82836	792054	82836

Note: This table presents estimates from Poisson and OLS models evaluating the overall impact of book digitization on loans and sales (Panel A), and the heterogeneous impact of book digitization on loans and sales for popular books as compared to the rest (Panel B). In columns (1) and (3) of both panels, the sample includes a balanced panel of 88,006 books over 9 years (2003-2011) for a total of 792,054 observations, while in columns (2) and (4), the sample includes data on 9,204 books over the same time period for a total of 82,836 observations. Loans represents the total number of times a book has been loaned in a given year within the Harvard system. Sales is the number of sold copies of that title in a year. Any-Loans and Any-Sales are indicator variables=0/1 depending on whether a book has been loaned or sold at least once in a given year, respectively. Post-Scanned equals one in years after a book has been digitized. Popular equals one for books that have more than one loan before the digitization program started (i.e., in 2003 and 2004). Book and year fixed effects are included in all models. Standard errors are in parentheses, clustered at the book level.

Table 3. **Robustness Checks**

Panel A: Alternate Specifications

	OLS		Log OLS		Poisson, no outliers	
	(1) Loans	(2) Sales	(3) Ln(Loans)	(4) Ln(Sales)	(5) Loans	(6) Sales
Post-Scanned	-0.0918	115.5	-0.0445	0.0421	-0.457	0.146
	(0.00325)	(69.32)	(0.00108)	(0.0126)	(0.0175)	(0.0827)
Book FE	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
N	792054	82836	792054	82836	791982	82764

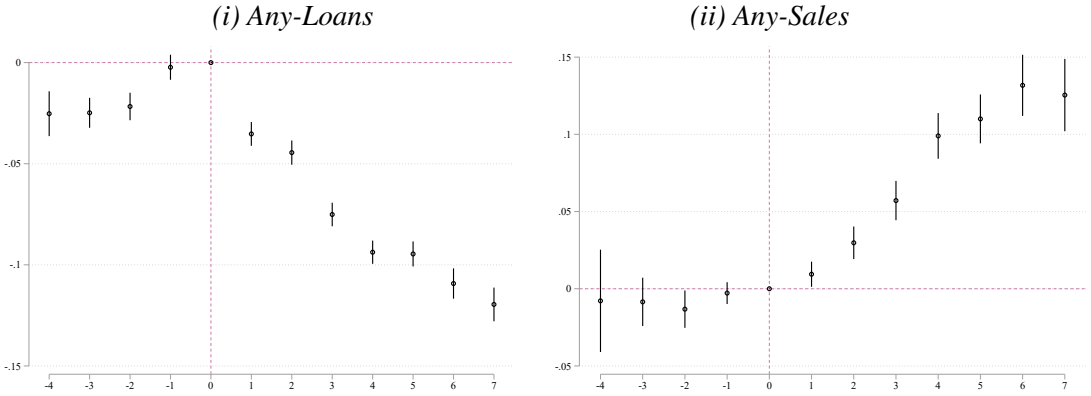
Panel B: Accounting for Editions

	Direct Effect		Accounting for Editions			
	(1) Editions	(2) Any-Editions	(3) Loans	(4) Sales	(5) Loans	(6) Sales
Post-Scanned	0.538	0.186	-0.478	0.225	-0.473	0.299
	(0.0371)	(0.00174)	(0.0150)	(0.131)	(0.0151)	(0.160)
Editions			-0.00517	0.00789		
			(0.00155)	(0.00239)		
Edition Grp. FE	–	–	–	–	Yes	Yes
Book FE	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
N	222003	792054	792054	26586	792054	26586

Note: This table presents the robustness of the baseline specification. Panel A evaluates robustness to alternate specifications and sensitivity to dropping outliers, while Panel B investigates the impact of digitization on the release of new book editions and the role of editions in driving the main effect. Loans represents the total number of times a book has been loaned in a given year. Sales is the number of sold copies of that title in a year. Post-Scanned equals one for years after a book has been digitized. In Panel A, the first two columns provide OLS estimates, the next two columns provide zero-inflated Log-OLS estimates (i.e., the dependent variable is $Ln(Loans_{it}+1)$ or $Ln(Sales_{it}+1)$), and the last two columns present Poisson estimates when dropping all books with more than 900,000 lifetime sales. In Panel B, estimates are presented from Poisson models, except column (2), which shows a linear probability model. Editions represents the number of editions available in print. Any-Editions is an indicator variable=1 if an edition is available at all in a given year. Edition Grp. FE includes ten fixed effects for the number of editions (starting at zero, with a common FE for all books with 8+ editions). See text for more details. Book and year fixed effects are included in all models. Standard errors in parentheses, clustered at the book level.

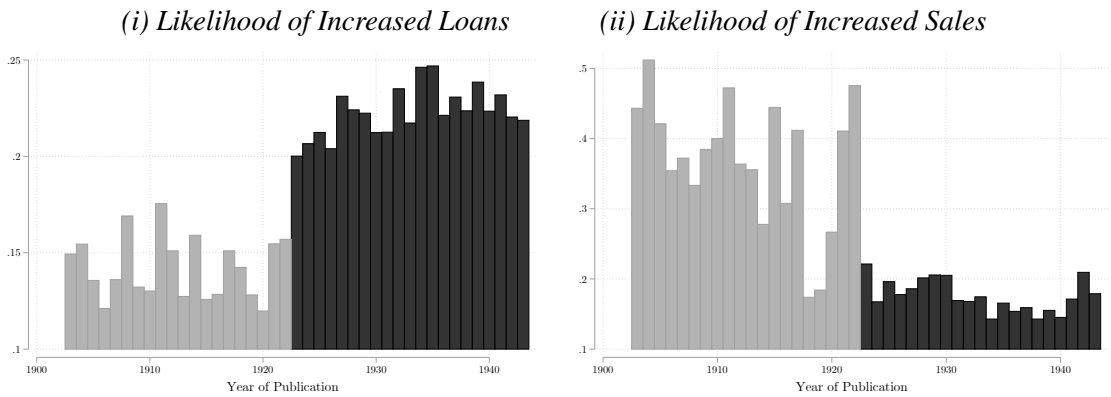
Figures

Figure 1. **Time-Varying Estimates of the Impact of Digitization**



Note: This figure provides visual illustrations of the event study specification: $Y_{it} = \alpha + \Sigma_z \beta_z (scanned)_i \times 1(z) + \gamma_i + \mu_t + \varepsilon_{it}$, where γ_i and μ_t represent book and time fixed effects, respectively, for book i and year t, $(scanned)_i$ equals one for all books that were eventually scanned and z represents the "lag," or the number of years that have elapsed since a book was first digitized ($= 0$ in the year before digitization. The main dependent variables here are Any-Loans (A) or Any-Sales (B). The chart plots values of β_z for different values of z. See section 4.2 for more details.

Figure 2. **Comparing Change in Demand for Pre-1923 and Post-1923 Books**



Note: This figure explores the impact of the digitization program on a cross-sectional sample of English language books originally published between 1904-1942, of which only those published before 1923 are digitized due to copyright restrictions. This includes 39,949 books with loans data, and 7,033 books with sales data. To construct this panel, we calculate the change in the number of loans and sales for a given book in the 2010-2011 period (after digitization) as compared to the 2003-2004 period (before digitization). We then plot the share of books in a certain publication year that increase their loans (or sales) on the y-axis and the publication year itself on the x-axis. Books published after 1923 (and which were not scanned) are indicated in black, and those before are indicated in gray.

Online Appendix

A Conceptual Framework

We introduce a simple motivating model that describes the consumer's decision to obtain the analog product with and without a free digital provider. As we show below, whether the arrival of a digital provider increases or decreases analog demand depends on two parameters: the search costs of finding a particular book, and the individual value from digital products.

Let b denote the book (which identifies its popularity), and let $s \in \{a,d\}$ be the seller (analog or digital, respectively). Consumer i's utility from buying book b through seller s is given as

$$u^i_{bs} = V^i_{bs} - c^i_b,$$

where V^i_{bs} is the book-specific monetary value, that is, the utility the reader gets from obtaining the book less its price. For any book, the analog value V_{ba} is fixed across consumers, but the digital value $V_{bd} \sim f[0,\bar{V}]$. Some consumers strongly value digital consumption (either due to taste or low cost), while others do not (perhaps because they have an aversion to digital copies or face transaction costs).

The search cost c^i_b depends on the book's popularity. For example, *Wealth of Nations* is well known, so that search costs are low for most consumers, whereas consumers may only find out about other titles through (costly) search. Across all consumers, the book-specific search cost is represented by a distribution $c_b \sim f[0,B]$, where the average search cost for less popular books is larger than that for more popular ones. The introduction of the digital provider decreases search costs to zero for all books, markets, and consumers because the digital provider introduces well-developed institutions for discovery.[19]

The consumer's decision

We distinguish between the utilities obtained in three cases: buying from an analog seller before digitization, buying from the analog seller after digitization, and buying from the digital seller after digitization.

Consumer i's utility from an analog seller when there is no digital provider can be written as

$$u^{i,pre}_{ba} = V_{ba} - c^i_b. \tag{3}$$

Because there is no digital option ($u^{i,pre}_{bd}$ is not defined), a consumer will buy the analog product if and only if $V_{ba} - c^i_b \geq 0$. After digitization, the search cost is eliminated. The consumer's utility from an analog seller is now

$$u^{post}_{ba} = V_{ba}, \tag{4}$$

[19] The intuition of the model holds even if search costs are lowered rather than eliminated.

and if the consumer were to choose the digital option, her utility would be

$$u_{bd}^{i,post} = V_{bd}^{i}. \tag{5}$$

With a digital option present, consumer i purchases the analog product if and only if the utility from doing so is larger than that from obtaining the free digital version, or $V_{ba} - V_{bd}^{i} \geq 0$.

The above equations suggest that the impact of free digital provision on analog demand depends on each consumer's search cost c_b^i and their valuation for the digital option V_{bd}^i. Figure D.1 illustrates the tension. For all consumers i with $c_b^i > V_{ba} > V_{bd}^i$, digitization enables the book's discovery because the previously high search cost is removed, and it leads to an analog sale that would not have otherwise happened. In contrast, for consumers with $V_{bd}^i > V_{ba} > c_b^i$, digitization did not lead to new discovery. Instead, the consumer substitutes the analog product for the digital provider. If both c_b^i and V_{bd}^i are larger (or smaller) than V_{ba}, the introduction of the digital provider will not change the consumer's decision to buy the analog version.

The *market-wide* impact of the digital provider on analog sales therefore depends on the distributions of search costs and of preferences for the digital option. If many consumers have high search costs, the discovery effect likely dominates and digital provision increases sales. But if search costs are generally low, for example with well-known books or when search institutions are otherwise readily available, the substitution effect may prevail and digitization likely cannibalizes analog demand.

B Additional Robustness Checks

B.1 Limiting the Sample

In the main analysis, we use all books that were checked out at least once between 2003 and 2011. The inclusion of all books could bias our results towards finding a positive effect of digitization on loans, if books were checked out more frequently late in our dataset due to digitization. To address this concern, we repeat our main analysis, including only books that were checked out at least once in the two years before the digitization period (2003/04). Table D.1 shows the coefficients of the baseline estimations for this sample. They are consistent with the main results in direction, size, and statistical significance.

B.2 Definition of Popularity

The main text shows that the impact of digitization varies by the title's popularity, with the most popular titles being relatively more negatively impacted. In the main text, we define a title as popular if it was checked out more than once before the digitization period (in 2003-04) – a cutoff that defines 12% of all books in our dataset as popular. Here, we examine the robustness of our results to this definition. Table D.2 replicates Panel B of Table 2, varying the popularity cutoff between zero and six loans before 2005 and

adding a popularity measure based on pre-digitization sales for the sales analysis. All results are consistent with those in the main analysis: digitization decreases loans of all books, but more so for popular works. It increases sales of less popular works, whereas its impact on sales of more popular works is generally positive, but smaller.

B.3 Other Mechanisms: Prices

We have found that digitization through Google Books leads to an increase in the number of editions as well as in the units sold per book, but the improved availability itself does not explain the increase in sales. In addition to the mechanisms studied in the main text, the increase in unit sales might also be driven by a decrease in prices, perhaps due to increased competition from the Google Books version and the new editions. If such decreases in prices are large enough, then digitization could lead to decreases in revenues.

While we do not have price information in the sales sample, we do have some price information from Bowker's Books-in-Print sample, which allows us to test whether and how digitization impacted at what prices new editions were introduced. Here, we treat each newly published edition as an observation, and we estimate the edition's suggested retail price as a function of the title's digitization status at the time of the edition's publication, in addition to indicator variables for the edition's year of publication and for each title. Formally, we estimate

$$Y_{jit} = \alpha + \beta PostScanned_{jit} + \gamma_i + \mu_t + \varepsilon_{jit}, \tag{6}$$

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

Table D.3 depicts the results from these regressions. They suggest that Google's digitization program had at most a small impact on prices of new editions. While the OLS regressions in columns 1-2 show no significant effect on the absolute prices of new editions, the log-OLS regressions in columns 3-4 provide some additional information. Column 3 finds that digitization decreases an edition's price by about 1.7% (p-value $= 0.08$). However, the coefficient decreases and loses statistical significance as we control for the number of available editions. That is, digitization impacts edition prices mostly through the number of available editions. Because the availability of editions does not fully explain the positive impact of digitization on sales (panel B of Table 3), neither would a decrease in prices.

C Regression Discontinuity

The main analysis takes advantage of variation in the timing and status of digitization across all books in the Harvard Widener library system. An underlying assumption in these analyses is that books that are digitized

are inherently similar to books that are not, or not yet, digitized. However, whether a book is digitized *at all* is a function of the book's copyright status. Throughout the time period of our study, all works that were originally published before 1923 are in the public domain and hence digitized, whereas works from 1923 and later were still protected by copyright and hence *not* digitized.

This discontinuity in copyright status is due to the most recent copyright extension in the United States. The 1998 Copyright Term Extension Act retroactively extended the copyright term for all protected works by twenty years, from 75 to 95 years for the works in our dataset. The copyright extension provides a sharp, exogenous discontinuity in the ex-post digitization status for works originally published around 1923. Beyond the digitization status, however, it is likely that the works from around that year are quite similar. Thus, were it not for the digitization of the older works between 2005 and 2009, one might expect analog demand for these titles to evolve similarly for works originally published on both sides of the 1923 cutoff.

Although the books are otherwise similar, there is a large discontinuity in how their demand has evolved between 2003/04 (before any works were digitized) and 2010/11 (when the digitization period at Harvard had ended). Figure 2 in the main text plots the difference in demand between the pre- and post-digitization periods for each year around 1923. Consistent with theory and the main empirical results, the figure suggests that digitized works are much less likely to see an increase in library checkouts but more likely to see an increase in physical sales through other outlets. Here, we utilize the jump in digitization more formally by using a regression discontinuity design. Formally, we estimate regression equations of the form

$$Y_j = \alpha + \beta Digitized_j + k(year_j) + \varepsilon_j, \tag{7}$$

where Y_j describes various measures of the change in book j's demand (library checkouts or sales) from 2003/04 to 2010/11, including the absolute unit change, an asymptotic sine transformation of these changes, and an indicator that is one if there is an increase in book j's demand.[20] Moreover, $Digitized_j$ is an indicator variable that is 1 if the book was digitized, a deterministic function of the book's original year of publication. We define $k(year_j)$ as a quadratic function of the book's publication year, centered around 1923, noting that lower- and higher-order polynomials provide very similar results. The bandwidth in each specification is its mean-squared-error optimal bandwidth.

Table D.4 shows the results from these specifications. The first three columns show results for changes in the Harvard library checkouts, and the last three columns focus on changes in sales. All results are supportive of those in the main text: treatment through digitization leads to a statistically significant decrease in the number of library checkouts, and an increase in the number of sales through outside channels. These results are robust to different bandwidths and functional forms of the publication year. Figure D.2 further illustrates the results, showing again that books originally published before 1923 and therefore digitized by Google Books are significantly less likely to see an increase in loans and significantly more likely to experience an increase in sales.

[20] We use an asymptotic sine transformation instead of the more common log-transformation because one would naturally expect many negative changes in demand, and dropping these may bias results.

D Appendix Tables and Figures

Table D.1. **Baseline Estimates Using Subset of Books With At Least One Loan Before 2005**

	Poisson		OLS	
	(1) Loans	(2) Sales	(3) Any-Loans	(4) Any-Sales
Post-Scanned	-0.681	0.165	-0.0252	0.0834
	(0.0287)	(0.103)	(0.00197)	(0.00605)
Book FE	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes
N	287523	53388	287523	53388

Note: This table presents estimates from Poisson and OLS models evaluating the overall impact of book digitization on loans and sales. The estimations are analogous to Panel A of Table 2 in the main text. However, they only include books that were checked out at least once pre-digitization (2003/04). This provides a balanced panel of 31,947 books in the loans estimations (columns 1 and 3), and of 5,932 books in the sales regressions (columns 2 and 4). Loans represents the total number of times a book has been loaned in a given year within the Harvard system. Sales is the number of sold copies of that title in a year. Any-Loans and Any-Sales are indicator variables=0/1 depending on whether a book has been loaned or sold at least once in a given year, respectively. Post-Scanned equals one in years after a book has been digitized. Book and year fixed effects are included in all models. Standard errors are in parentheses, clustered at the book level.

Table D.2. **Estimates by Popularity: Different Cutoffs**

Panel A. Impact on Loans

	Popularity Cutoff: Pre-2005 Loans			
	>0	>2	>4	>6
Post-Scanned	-0.119	-0.419	-0.445	-0.454
	(0.0339)	(0.0133)	(0.0130)	(0.0151)
Post Scanned x Popular	-0.933	-0.458	-0.409	-0.225
	(0.0983)	(0.182)	(0.229)	(0.282)
Book FE	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes
N	792054	792054	792054	792054

Panel B. Impact on Sales

	Popularity Cutoff: Pre-2005 Loans				Pre-2005 Sales
	>0	>2	>4	>6	>0
Post-Scanned	0.401	0.341	0.305	0.297	0.994
	(0.211)	(0.187)	(0.154)	(0.154)	(0.373)
Post Scanned x Popular	-0.278	-0.185	-0.628	-0.124	-0.701
	(0.228)	(0.223)	(0.157)	(0.217)	(0.399)
Book FE	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes
N	82836	82836	82836	82836	82836

Note: This table presents estimates from Poisson models evaluating the heterogeneous impact of book digitization on loans (Panel A) and sales (Panel B) for popular books as compared to the rest. In Panel A (loans), the sample includes a balanced panel of 88,006 books over 9 years (2003-2011) for a total of 792,054 observations; in Panel B (sales), the sample includes data on 9,204 books over the same time period for a total of 82,836 observations. Loans represents the total number of times a book has been loaned in a given year within the Harvard system. Sales is the number of sold copies of that title in a year. Post-Scanned equals one in years after a book has been digitized. Popularity cutoffs are according to the number of library loans before the digitization program started (i.e., in 2003 and 2004). One exception, in Panel B, Column 5, the cutoff is based on sales in 2003 and 2004. Book and year fixed effects are included in all models. Standard errors are in parentheses, clustered at the book level.

Table D.3. **Impact of Digitization on Prices of New Editions**

	OLS		Log OLS	
	(1) Price	(2) Price	(3) Ln(Price)	(4) Ln(Price)
Post-Scanned	-0.00438	0.106	-0.0174	-0.0143
	(0.775)	(0.777)	(0.00996)	(0.00998)
Editions		0.0220		0.000638
		(0.00585)		(0.000110)
Book FE	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes
N	280240	280240	280115	280115

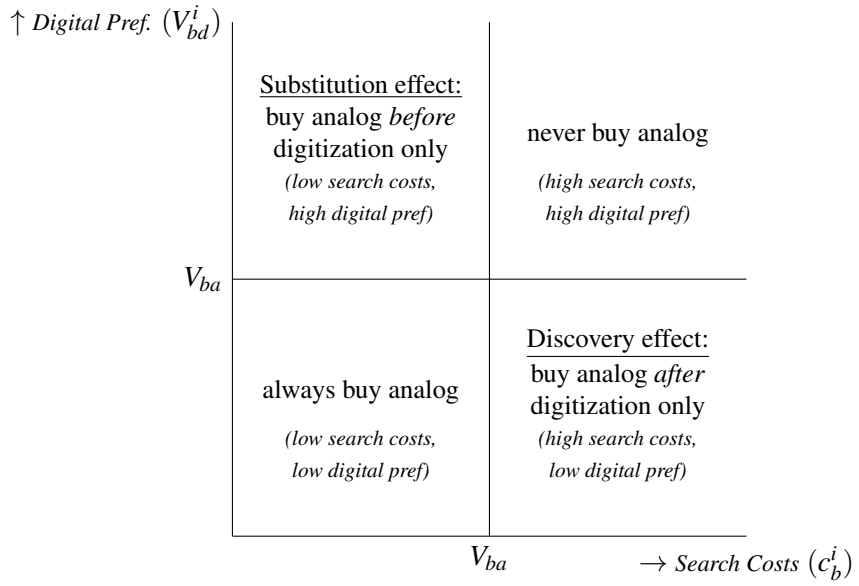
Note: This table presents estimates from OLS and log-OLS models evaluating the overall impact of book digitization on prices of new editions. This estimation is on the edition-level, for the 280,115 editions published between 2003 and 2011 of the 24,667 titles we matched between Harvard's library system and the Bowker Books-in-Print directory. Post-Scanned equals one in years after a book has been digitized. Book and year indicators are included in all models. Standard errors are in parentheses, clustered at the book level.

Table D.4. **Regression Discontinuity Estimates**

	Loans			Sales		
	(1) Loans	(2) Asinh(Loans)	(3) Increase	(4) Sales	(5) Asinh(Sales)	(6) Increase
Digitized	-0.219	-0.144	-0.0641	577.4	0.277	0.159
	(0.0297)	(0.0175)	(0.00781)	(450.8)	(0.126)	(0.0229)
N	47902	47902	47902	8016	8016	8016

Note: This table presents results from regression discontinuity estimations. The dependent variables are functions of the changes in analog demand (loans and sales) between 2003/04 (before digitization) and 2010/11 (after digitization). In columns 1 and 4, it is the absolute change in analog demand; columns 2 and 5 use the asymptotic sine of that change; and columns 3 and 6 use an indicator that is 1 if analog demand has increased. The independent variables of interest, *Digitized*, is an indicator that is 1 if the book's original year of publication is before 1923. A quadratic function of the publication year is included. The bandwidth in each specification is the MSE-optimal bandwidth. Robust standard errors in parentheses.

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

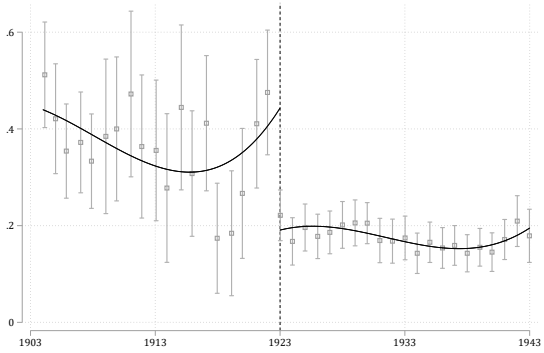


Note: This figure provides an illustration of predictions from the theoretical framework. The framework models an individual customer i's decision to purchase an analog version of the book (a physical copy) as a function of his or her search costs c_b^i (x-axis) and preference for digital copies V_{bd}^i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

Figure D.2. **Annual Estimates of Regression Discontinuity**

(i) Estimated Likelihood of Increased Loans

(ii) Estmated Likelihood of Increased Sales



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