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

February 21, 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. Shane Greenstein, Aruna Ranganathan, Joel Waldfogel, Pam Samuelson, and attendees of the SERCI Congress, Toronto 2018 and Toulouse Digital Economics Conference 2019 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 catalog of existing works. Much like a modern-day Library of Alexandria, there is the real possibility that the Internet could serve as a repository of all knowledge in digital form (Samuelson, 2011). Increasing access to past knowledge could have massive implications for 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. 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. This result can be largely attributed to legal considerations, especially copyright challenges from traditional publishers and authors that have been litigated in the US Supreme Court.[1] 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.[2] 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.[3] If the negative effects of digitization of physical works on the demand for print are indeed low, digitization might be a win-win for consumers, publishers, and authors. In this paper, we move beyond the theoretical debates and attempt to provide empirical estimates of whether and to what extent digitization harms the consumption of physical works.

We begin by clarifying the legal debates via a simple conceptual framework. On one hand, digitization can lead consumers to substitute digital alternatives, cannibalizing demand for physical works. On the other, digitization could also lower search costs and increase discovery for obscure works, stimulating demand

[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] 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/.

[3] 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. See Jsevold (2016).

for physical works. Our quantitative analysis, which is the heart of our paper, sheds light on this tension between discovery and substitution.

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 a modern-day Library of Alexandria. The project was launched in partnership with Harvard University's Widener library (as well as a few others), which provided public domain books and texts to be digitized by Google. Given the magnitude of the effort at Harvard, the process of scanning works and making them available online continued from 2005 till at least 2009. We were able to obtain proprietary data from Harvard on over half a million digitized and non-digitized works, including the date on which a particular work was digitized (if applicable). While the order of digitization of books was not deliberately randomized, it was not explicitly selective and did not prioritize the most interesting books for early digitization.[4] We exploit the variation in the timing of 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-difference analysis with book and year fixed effects. In a parallel set of analyses, we rely on the fact that US books published after 1923 were not digitized (because their copyright status was not clear), while those before were. We exploit the sharp discontinuity in digitization status across this cutoff for additional estimations.

We combined data from three sources for our analysis. First, we collected data on library loans within the Harvard system for about 90,000 books with at least one loan between 2003 and 2011. Second, for a subset of 9,204 books within this sample, we obtained weekly US sales data on all related editions from the NPD (formerly Nielsen) BookScan database.[5] Finally, we also collected data on publications of new editions from the Bowker BooksInPrint database to examine the effect of digitization on bringing out-of-print books back to life.

Armed with these matched data, we compare the loans and sales for digitized books to non-digitized books before and after the digitization year in a difference-in-differences setting. We find that the impact of digitization on use is negative when considering internal readership 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 35%. Alternatively, digitization reduces the probability that a book will be loaned at least once in a year by about 6% but increases the likelihood of at least one sale by 8%. We further find that the positive effect of digitization on sales is largely driven by less popular books. These

[4] Rather, our fieldwork suggests that digitization proceeded on a "shelf-by-shelf" basis and was driven by convenience.

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

findings are in line with our conceptual framework that digitization can increase sales through discovery but could cannibalize demand when search costs are low. Accordingly, our findings indicate that the effects of digitization on demand are negative for popular books (which are well-known) and for readers within the Harvard system (where trained librarians facilitated search). 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 explains only a small part 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. Copyright holders' concerns about the possible loss of revenue from digital availability have proved to be an important roadblock to making digitized works available online. For example, in Author's Guild vs. Google (2013), the plaintiffs argued that Google Books will negatively impact the market for books by serving as a "market replacement." Google contended that their digitization project would boost sales by making it easier for readers to find books. Our study helps move beyond these theoretical debates. In particular, our finding that digitization, on the whole, increased sales suggests that concerns about market replacement effects are likely overblown. In contrast, copyright holders might be able to increase demand through digitization, especially for less popular and out-of-print works.

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 the price of books (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, 2018) 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 important policy debate around the mass digitization of books and examining the impacts of weakening copyright restrictions through digitization.

2 Background and Conceptual Framework

2.1 The Google Books Project: A Brief Background

The Google Books project (originally known as the Google Print Library Project)[6] 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. As soon as these works were scanned, they were usually made digitally available on the Google Books website for the general public to read.

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] The lawsuits were centered on the idea that Google's digitization effort caused material harm to the authors and the publishers of the printed works, violating their exclusive rights to profit from their works under the terms of copyright law.[8] 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.[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] Empirical evidence on either side on the realized effects of digitization 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 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.

2.2 Conceptual Framework

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

[6] 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] See https://tinyurl.com/y7hsxalw.

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

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

searchability, and the substitution effect of digital distribution as a competitor for existing, physical products. To clarify this theoretical tension, we develop a simple conceptual framework that we describe here and depict visually in Figure 1. At the heart of this framework lies the idea that digitization – as a repository of all knowledge – can lower search costs for otherwise obscure works.[11] A more formal exposition of this framework is in Appendix A.

Before digitization, all consumers who value the book more than its search cost will purchase the physical book. After digitization, search costs are no longer relevant. Instead, consumers now face a choice between a digital and an analog option. All consumers face zero search costs, but those who value the analog book more than the digital option will purchase the physical book. Therefore, whether demand for the physical option increases or decreases after digitization depends on two parallel effects. First, how many consumers who originally purchased the analog copy will now move to the digital copy? We call this the substitution channel. And second, how many new consumers will purchase a physical copy given that the search costs have disappeared? We call this the discovery channel.

These two channels are depicted in Figure 1. For a given book, consumers are divided into four quadrants based on whether they face high search costs (x-axis) and whether they exhibit a taste for the digital alternative (y-axis). The substitution effect comes from consumers in the top left quadrant who found it valuable to consume the analog version pre-digitization given their low search costs but now abandon physical copies given their preference for the digital alternative. The discovery effect comes from consumers in the bottom right quadrant who faced high search costs and therefore didn't consume pre-digitization but now discover the book given zero search cost and consume the analog version due to their low digital preference. The diagonal elements are irrelevant to the net effect of digitization on change in demand for physical versions of books. The net effect simply depends on the relative mass of consumers in the two off-diagonal quadrants.

Our framework highlights the importance of empirical analysis to estimate the market-wide effect of digitization. In our empirical analysis, we estimate this market-wide effect for the entire set of books in our sample. Further, the substitution and discovery effects are likely to be book- and consumer-specific. For example, consider a very popular book (such as Jane Austen's *Pride and Prejudice*) for which search costs are virtually zero. For such books, the discovery effect is minimal, and only the substitution effect applies, leading to a decrease in demand. In contrast, consider a relatively obscure book that had virtually zero consumption before digitization. For such books, the discovery effect likely dominates, leading to an increase

[11] Before digitization, a reader might not know about a book or the book might be difficult to locate. After digitization, all books can be easily located through a simple search on the Google Books website.

in demand. Similarly, consider variation in search costs across consumers. Consumers who belong to an institution that facilitates search (such as Harvard) are more likely to be affected by the substitution effect rather than the discovery effect. In addition to estimating the market-wide effects, our empirical examination also examines the effect on this variation across books and consumers.

3 Data and Research Design

3.1 Google Books and Harvard Libraries' Natural Experiment

To empirically evaluate the effect of digitization on the demand for physical works, one needs variation in digitization across a wide variety of books linked to data on demand for the physical alternatives. We focus our study on the digitization of works from Harvard's libraries through the Google Books project to make progress on this question.

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 that were already in the public domain and for which the copyright was deemed to have expired. This included public domain works from Harvard's largest and most prestigious Widener Library that houses a total of over 3.5 million books in its collections. For example, works published in the United States before the year 1923 were provided for scanning given that all works published before that date have fallen in the public domain. Due to changes in copyright law, including the Copyright Term Extension Act of 1998 ("Mickey Mouse" law), this date would not change until much after the digitization program was completed.

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. 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. 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.

3.2 Data

The underlying database contains the entire record of the Harvard Libraries holdings that were scanned, as well as all works published between 1923 and 1943 that were not scanned. For these books, we collected data from three sources. First, we obtained proprietary data on all library checkouts between 2003 and 2011. These data contain information on the specific patron code checking out the work, including whether and when a book was checked out by Google Books. Because books were usually made available a few weeks after digitization, this date of digitization approximates the date when a book was made available online.

Second, we obtained access to NPD (formerly Nielsen) BookScan, which provides weekly sales information for printed books. Nielsen (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.[12] Because our data from Harvard do not contain global unique identifiers (i.e., ISBNs), we manually searched NPD BookScan for the book titles to find suitable matches, aggregating sales of all editions for each title. Given the tedious data collection process, we searched for sales data for 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 collected 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 matched titles in the Harvard database to this database and were able to find matches for almost 25,000 unique titles. 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

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

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 about 88,000 books (those with at least one loan), 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 our baseline specification, we exploit this variation across years in the timing of the digitization, as well as the fact that many books were never digitized. Further, we have 2,954 books with at least one sale, and 24,667 books with at least one edition in the Bowker Books-in-Print database. 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 of these 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

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

where $PostScanned_{it}$ is an indicator that is 1 if book i has been made available on Google Books before year t, and γ_i and μ_t are book and year fixed effects, respectively. The dependent variable, Y_{it}, denotes book- and year-specific measures of demand (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 in our context with skewed dependent variables (Wooldridge, 1999). In a parallel set of 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

We first estimate the impact of digitization on demand through traditional channels: library checkouts (through Harvard's Widener library), and sales of physical copies. Table 2 Panel A displays the results from this specification using 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 the number of loans through 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. 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). 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

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

where γ_i and δ_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.[13] We estimate this equation corresponding to the linear probability models estimated in Table 2.

[13] 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.

Panel A of Figure 2 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, but 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 popularities. If search costs are otherwise high (for example, for obscure books), then the discovery effect of Google Books should outweigh the substitution effect. We repeat the analyses from Table 2, with an additional interaction term of the Post-Scanned variable with an indicator that equals 1 if the book was checked out at Harvard's libraries more than five times in the three years before digitization. Most books are rarely checked out, so we assume that the interest within Harvard is a proxy for the book's popularity.[14]

Table 2 Panel B shows the results from this exercise. Consistent with expectations, the impacts of digitization vary widely for both loans and sales, as more popular books are affected more negatively. Following Columns 1 and 3, all digitized books see a significant decrease in loans compared to books that were not digitized or digitized later. However, the impact is significantly larger for popular works, which experience a decrease of about 52% ($= e^{-0.447-0.303} - 1$) compared to a decrease of 36% for less popular books, according to the point estimates in column 1. Alternatively, while the likelihood that a book will be checked out reduces by about 5.9 percentage points for less popular books, for more popular books this estimate is almost 36 percent points. The relative impacts on sales between popular and less popular books are similar when considering the Poisson estimates. The more obscure books experience an almost 40% increase in sales – consistent with facilitated discovery outweighing substitution – whereas the estimated impact on sales of popular books is negative and of the order of 10%. In other words, publishers' and authors' concerns about the cannibalization of their work via digital substitution is real, but only for a small set of popular books. For

[14]This definition identifies 1,232 books as popular, accounting for 1.4% of books in the loans data, and 9.5% of books in the sales data.

the vast majority of obscure books, the positive effect on demand through discovery outweighs the negative effect of substitution. This picture changes somewhat when considering the likelihood of making at least one sale. Here, both popular and less popular books benefit from digitization. In fact, it seems like a larger share of popular books benefit from digitization in terms of making at least one sale. However, 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 1,300 copies a year, and over 11,600 copies over our sample period, according to our data.

4.4 Robustness Checks

While the above results are consistent with our conceptual framework, one might be concerned that they are consequences of our modeling choices or of large, endogenous differences across the treated and control groups. To address these concerns, we perform three sets of robustness checks, including estimating alternate specifications, accounting for the role of new editions in driving increased demand and using the sharp cutoff around 1923 to provide an alternate way to assess the impact of digitization.

Alternate Specifications: First, we use OLS, with the dependent variable being either Y_{it}, or $Ln(Y_{it}+1)$. Estimates from these specifications are presented in Table 3, Panel A. The first two columns present results from OLS models and the next two present those from Log models. 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.

Accounting for Editions: Next, we consider the possibility that the freely available Google Books version allows other publishers to create and publish more and higher-quality copies of the text, in turn allowing consumers to buy editions that did not exist previously. Anecdotally, Google Books reduced the barrier to entry for publishers who could find a way to download entire scanned versions of books and make them available in print. We examine this possibility here. We first estimate the impact of digitization through Google Books on the number of *new* editions that enter the market, and we then examine whether the changes in use and demand can be attributed to improved availability.

Panel B of Table 3 shows the results of these specifications estimated using Poisson models similar to the baseline analysis. The first column shows the first stage: the impact of digitization on the number of newly available editions. Titles become available in many more editions after their digitization, with an average increase of 71% ($= e^{.538} - 1$).

The remaining columns of Table 3 Panel B estimate the impacts of digitization on use, controlling for the number of newly introduced editions of the work.[15] Columns 2 and 3 control for new editions linearly, while columns 4–7 provide more flexible controls, including dummy variables for each number of new editions, combining observations with more than eight editions in one group.[16] Estimates from all regressions show that changes in access – while present – do not fully explain the previously estimated effects. While new editions impact demand, the direct impacts of digitization on loans and sales remain statistically significant and of similar sizes.

Exploiting the discontinuity around 1923: Finally, while our analysis so far relies on the timing variation in the digitization of books across Harvard, there is another source of variation that can be exploited 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 digitizated 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 across the two periods increased. The average value of this indicator by publication year is presented in Figure 2, Panel B 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 change in the increase in loans and sales around 1923. Loans are less likely and sales are more likely to increase for digitized works, and the difference is quite sharp around 1923. Further, the darker and the lighter bars are relatively flat within a given period, 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 and confound the causal interpretation of our estimates.

[15] We treat all titles without a match in the Bowker Books-in-Print database as out-of-print titles. Some of these may be false negatives, however. To account for this, we repeat the analyses, dropping all non-matches. The results are almost unchanged.

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

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 Panel B is quite reassuring. In Appendix B, 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

The fear that digital availability will cannibalize the use of printed works and cause financial harm has largely blocked projects that have tried to create a centralized and digital repository containing all books ever published. In this paper, we provide empirical evidence on the relationship between digitization and the demand for physical works in the context of the digitization of books from Harvard's Widener Library by Google Books. Our analysis suggests that the discovery effect outweighs the substitution effect for the majority of works digitized by Google Books, especially when institutions for discovery are not otherwise in place or when books are relatively unpopular.

Our results have 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 copyright holders in terms of the use of printed works. Combined with evidence from other studies in this literature that find no harm of digitization to consumers (e.g., Chen et al., 2018), our estimates suggest that existing copyright holders should be more supportive of the digitization of their catalog. 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.

Second, 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 books from a single, albeit important, library's collections, and our evidence is restricted to books in the public domain. It is possible that these effects could be different for a more general sample of contemporary books. Further, 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. The age of digitization provides a real opportunity to freely provide easily accessible repositories of the sum of human knowledge. Our empirical evidence suggests that fears of cannibalization might be overblown. As such, 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.

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*.

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).

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

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. (2018, February). Copyright and Generic Entry in Book Publishing. SSRN Scholarly Paper ID 2938072, Social Science Research Network, Rochester, NY.

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.

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.

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)	0.01	0.12	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 five loans before the digitization program started (i.e., between 2003 and 2005). 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	-0.447	0.335	-0.0593	0.0701
	(0.0129)	(0.159)	(0.00167)	(0.00497)
Post-Scanned x Popular	-0.303	-0.448	-0.308	0.0817
	(0.235)	(0.158)	(0.0130)	(0.0162)
Book FE	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes
N	792054	82836	792054	82836

Notes: 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 five loans before the digitization program started (i.e., between 2003 and 2005). 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 Models	
	(1) Loans	(2) Sales	(3) Ln(Loans)	(4) Ln(Sales)
Post-Scanned	-0.0918	115.5	-0.0445	0.0421
	(0.00325)	(69.32)	(0.00108)	(0.0126)
Book FE	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes
Popularity Effect	No	No	No	No
N	792054	82836	792054	82836

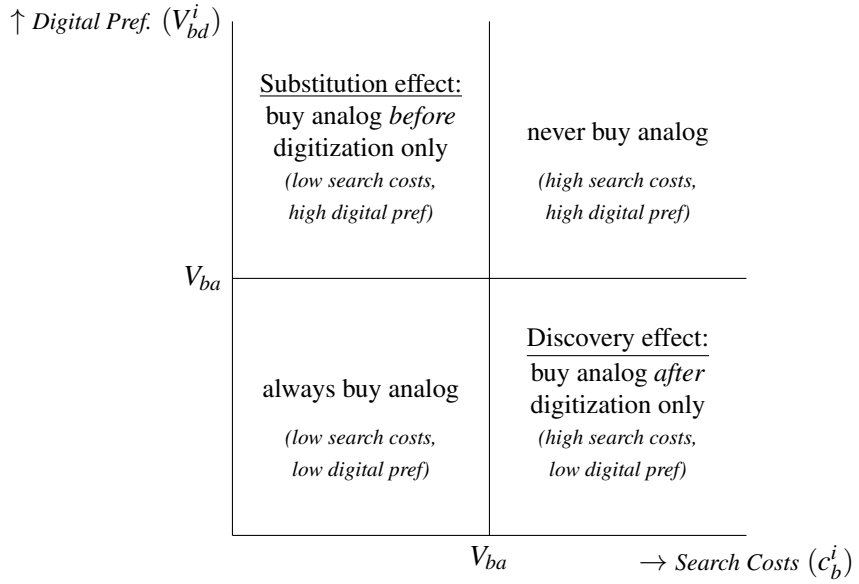
Panel B: Accounting for Editions

	Direct Effect	Accounting for Editions					
	(1) Editions	(2) Loans	(3) Sales	(4) Loans	(5) Sales	(6) Any-Loans	(7) Any-Sales
Post-Scanned	0.538	-0.478	0.225	-0.473	0.299	-0.0618	0.0545
	(0.0371)	(0.0150)	(0.131)	(0.0151)	(0.160)	(0.00170)	(0.00493)
Editions		-0.00517	0.00789				
		(0.00155)	(0.00239)				
Edition Grp. FE	–	–	–	Yes	Yes	Yes	Yes
Book FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	222003	792054	26586	792054	26586	792054	82836

Notes: This table presents the robustness of the baseline specification. Panel A evaluates robustness to alternate specifications 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 and the next two columns provide zero-inflated Log-OLS estimates (i.e., the dependent variable is $Ln(Loans_{it}+1)$ or $Ln(Sales_{it}+1)$). In Panel B, estimates are presented from Poisson models, except columns (6) and (7), which rely on OLS specifications. Editions represents the number of editions available in print. Edition Grp. FE includes ten fixed effects for the number of editions (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. **Theoretical Framework: Decision to Consume Analog vs. Digital**

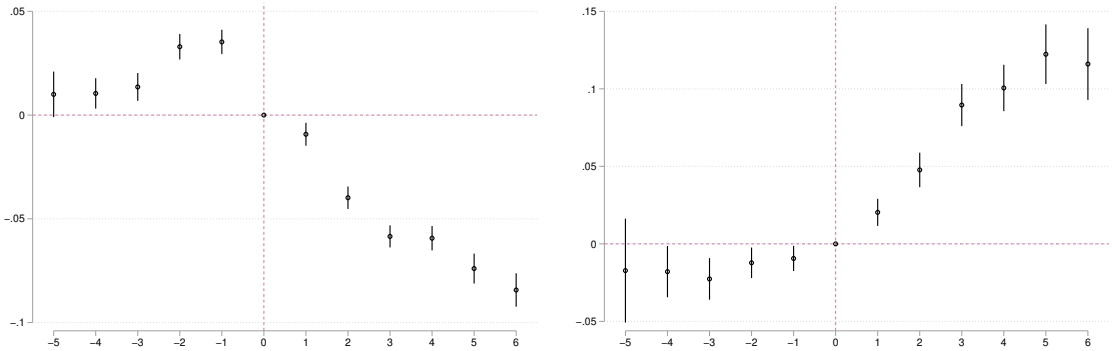


Note: This figure provides an illustration of predictions from the theoretical framework. The framework models an individual customer i's decision to purchase an analog version of the book (a physical copy) as a function of his or her search costs c_b^i (x-axis) and preference for digital copies V_{bd}^i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

Figure 2. **Graphical Evidence for the Impact of Digitization on Demand**

A. Panel Estimates for Years Before and After Digitization

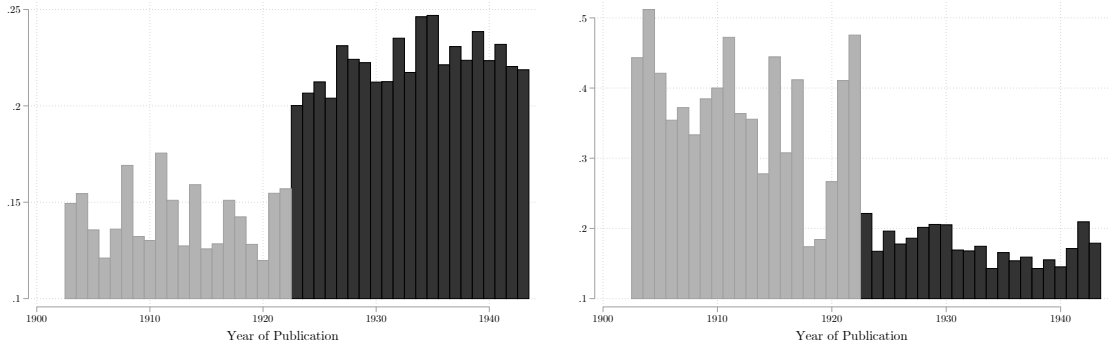
(i) Any-Loans *(ii) Any-Sales*



Note: This figure provides visual illustrations of the event study specification: $Y_{it} = \alpha + \Sigma_z \beta_z (scanned)_i \times 1(z) + \gamma_i + \delta_t + \varepsilon_{it}$, where γ_i and δ_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. 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.

Panel B. Comparing Pre-1923 to Post-1923 Books

(i) Likelihood of Increased Loans *(ii) Likelihood of Increased Sales*



Note: This figure explores the impact of the digitization program on a cross-sectional sample of English language books 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, *Pride and Prejudice* 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.

The consumer's decision

Given the structure of the utility function, 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

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

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 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, the substitution effect may prevail and digitization likely cannibalizes analog demand.

Note that this model also allows the impact to vary across customers with different costs of access to analog copies. If a consumer belongs to an institution with mechanisms to facilitate search (for example, university libraries), then the search costs c_b^i for all books are likely close to zero and the substitution effect likely prevails. On the other hand, if consumers do not have access to these institutions, then the tension between the discovery and substitution effects is more pressing.

B 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 during the time of this study, 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 years to 95 years for the works in our dataset. This policy change provides another source of variation, which we can use to estimate the impact of digitization through Google Books on the demand for a work. 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 Panel B 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

$$Y_j = \alpha \times Digitized_j + k(year_j) + \varepsilon_j, \tag{6}$$

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.[17] 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 B1 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 B1 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.

[17] 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.

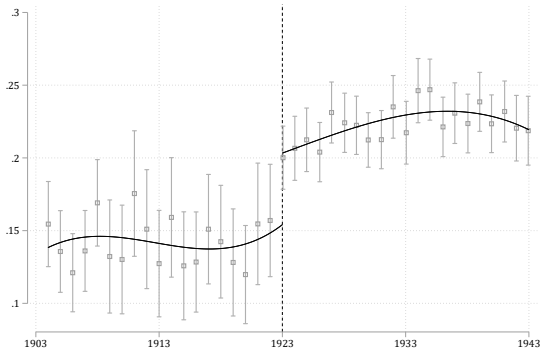
Table B1. Regression Discontinuity Estimates

	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

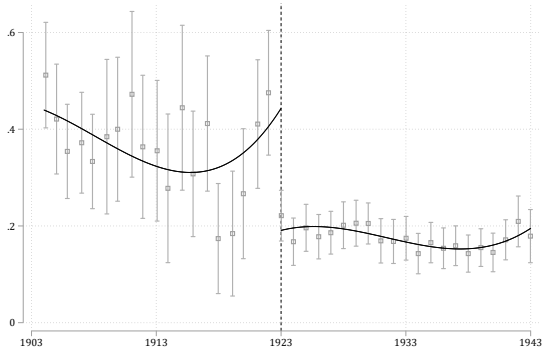
Notes: This table presents results from a regression discontinuity regression. The dependent variables are functions of the changes in analog demand (loans and sales) between 2003/04 (before digitization) and 2010/11 (after digitization). In column (1), it is the absolute change in analog demand; column (2) uses the asymptotic syne of that change; and column (3) uses an indicator that is 1 if analog demand has increased. The independent variable of interest, *Digitized*, is an indicator that is 1 if the book'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 B1. Annual Estimates of Regression Discontinuity

(i) Estimated Likelihood of Increased Loans



(ii) Estmated Likelihood of Increased Sales



Note: This figure presents estimates of the impact of the digitization program on a cross-sectional sample of English language books of which only those 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. 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.