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

January 9, 2019

Abstract

The age of digitization promised to deliver a centralized, digital repository of all knowledge. However, copyright holders' concerns about the cannibalization of physical works has prevented the realization of this vision. We shed light on this topic using newly collected data on timing of the digitization of books from Harvard's Widener Library by the Google Books project. While digitization does reduce loans of physical works within Harvard, it increases sales of physical editions to the tune of about 38%, especially for less popular works. These results suggest that, rather the cannibalizing demand, digitization might benefit copyright holders through increased discovery for 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 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). 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 by last count, over 25 million books had already been digitized through these efforts (Somers, 2017).

Despite the technological progress and the financial investment, the vision of a "digital Library of Alexandria" has not come to life. There 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] Authors and traditional publishers 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 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, clearing the way for the long-standing vision of providing access to the entire catalog of human knowledge in a single repository. 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.

First, following the legal debate, we argue that digitization can have both substitution and discovery effects. After digitization, consumers could substitute from physical to digital alternatives, thereby cannibalizing demand. On the other hand, digitization could lower search costs and increase the discovery of more obscure

[1] Despite court decisions in favor of Google's digitization project, the company "all but shut down its scanning operation" (Somers, 2017).

[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 Jøsevold (2016).

works, thereby stimulating demand for physical works. The relative importance of these two effects is likely to determine the overall effect of digitization on demand for physical works. We sketch a simple conceptual framework that clarifies this tradeoff.

Our quantitative analysis sheds light on this tension. We focus on the Google Books project which was launched in 2004 with a vision to digitize all works ever created and was the leading contender to be the 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, the process of scanning works and making them available online, started in 2005 and continued 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 this order was not deliberately randomized, the timing of the digitization of books was not explicitly selective 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 variation in the timing of digitized books as well as between books which were or were not scanned, to evaluate the impact of digitization on the demand for physical works.

We focus on two subsets of the broader data 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 2003-2011. Second, for a subset of about 9,200 books within this sample, we obtained weekly US sales data on all related editions from the NPD (formerly Nielsen) BookScan database.[4] 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 as compared to non-digitized books before and after the digitization year in a difference-in-differences setting controlling for non-parametric book and calendar year fixed effects. We find that the impact of digitization on use is negative when considering internal readership at Harvard, but is positive when considering sales. Specifically, digitization lowers the number of checkouts within Harvard by about 38%, but increases sales by 36%. We also find that the positive effect of digitization on sales is largely driven by less popular books, as digitization reduces sales for the smaller subset of popular books. These findings are in line with the theory that digitization can increase sales through discovery. However, this discovery effect is minimal when search costs are already quite low. We therefore find that the effects of digitization on demand are negative for

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

popular books (which are well-known) and for readers within the Harvard system (where the library's website and 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. While this "bringing books back to life" effect if quite important, it 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 by providing data driven and causal estimates of the impact of digitization on demand for printed works. In particular, our finding that digitization, on the whole, increased sales, suggests that concerns about market replacement effects are likely to be 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, 2018), 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 role of weakening copyright restrictions through digitization on demand for physical books.

Our paper proceeds as follows. In Section 2 we discuss the Google Books project and provide a theoretical framework for our empirical discussion, Section 3 describes our data and research design, Section 4 presents the quantitative results and Section 5 concludes with a discussion of the implications of our results for copyright and digitization policy.

2 Background and Theoretical Framework

2.1 The Google Books Project: A Brief Background

The Google Books project (originally known as the Google Print Library Project)[5] was announced by Google in December, 2004. The aim of the project was to make offline information contained in printed works available and searchable online. At the project's inception, Google partnered with the libraries of Harvard University, Stanford University, the University of Michigan and the University of Oxford as well as the New York Public library 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 a group of authors and publishers, including the Authors Guild and the Association of American Publishers, who filed class action suits against Google for copyright violation.[6] 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, and were illegal under the terms of copyright law.[7] 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.[8] The argument here was that Google Books' digitization efforts increased the discovery of printed works and "increase[d] the visibility of in and out of print books, and generate[d] book sales."[9] 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).

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

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

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

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

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

2.2 Conceptual Framework

The Google Books case thus depends on two counteracting forces: the discovery effect of Google Books as an aggregator of information that increases awareness of certain works, and the substitution effect of free digital distribution as a competitor for existing, physical products. To clarify this theoretical tension, 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

$$u_{bs}^{i} = V_{bs}^{i} - c_{b}^{i},$$

where V_{bs}^{i} is the book-specific monetary value – 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_{b}^{i} 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

$$u_{ba}^{i,pre} = V_{ba} - c_b^i. \tag{1}$$

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

$$u_{ba}^{post} = V_{ba}, \tag{2}$$

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

$$u_{bd}^{i,post} = V_{bd}^i. \tag{3}$$

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 – as might be the case 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, the tension between the discovery and substitution effects is more pressing.

3 Data and Research Design

3.1 Google Books and Harvard Libraries' Natural Experiment

While the copyright cases described above were quite contentious and involved a number of parties and issues, our focus in this paper is much narrower. We focus our study on the digitization of works from Harvard's libraries through the Google Books project. Given the potential for significant legal challenges to Google's efforts, Harvard Libraries' participation in the Google Books project was limited to works that were already in the public domain and for whom 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. Specifically, works published in the United States before the year 1923, and those published internationally before 1909 were provided to Google for scanning. We focus on the real effects of the digitization of these works in order to shed light on the broader implications of the digitization of printed works on readership and sales.

The digitization of Harvard's public domain books proceeded as follows. The Google Books project had 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 would be "loaned" to Google under this special code to be taken to the scanning facility. Once the book had been scanned, it would be returned to the library and also made available on the Google Books website after a short delay.

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, the order in which books were scanned was driven by convenience, rather than an explicit selection mechanism. Specifically, the 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. It is this quasi-random variation in the timing of the scanning project that we exploit to estimate the impact of digitization on eventual readership and sales. Figure **??** shows the number of books that were digitized in each year between 2005 and 2009.

3.2 Data

We obtained proprietary data from the Harvard Libraries with an entire record of their holdings that were scanned, as well as all works published between 1923 and 1943 that were not scanned. We also obtained separate information on all library checkouts between 2003 and 2011. These data contain information on the specific patron code checking out the work (for example, faculty, student or visitor), including whether a book was being checked out by Google Books. This allows us to estimate the time when a book was digitized and when it was made available online.

In addition to internal data from Harvard libraries on book digitization and loans, we also collected data from two other sources. First, we obtained access to NPD (formerly Nielsen) BookScan, which provides weekly sales information for printed books. These data are collected through tracking book sales using scanner data from a large panel of retail booksellers including major bookstore chains, discount retailers such as Costco and major online retailers like Amazon. They claim to track about 85 percent of total retail sales.[10] Our data from Harvard do not contain global unique identifiers such as ISBN numbers, so 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 Harvard collection before 1943 with at least 4 loans between 2003 and 2011, for a total of 9204 titles.

Second, we also 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 is 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. This is, to our knowledge, the first dataset that matches the digitization status of works with data on their sales and in-print status.

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 reuse and sales. We organize these 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 not digitized and 37,743 that were. We also have 2412 books with at least one sale, and 24,667 books with at least one edition in the Bowker Books-in-Print database. These data are summarized in Table 1. The average book has about 0.25 loans, sells about 1640 copies and adds 1.08 editions each year, although the median value for all three of these outcomes is zero.

[10] See (Sorensen and Rasmussen, 2004) and https://tinyurl.com/y94qpsqt, accessed June 26, 2018.

3.3 Preliminary Evidence

Our research design relies on the randomness of the timing of digitization, including whether a book is digitized at all. That is, we assume that – on average – digitized books would have followed a similar path as books that were not digitized, were it not for their digitization. We implicitly test this conjecture in Figure **??**, which plots the average annual loans (left panel) and sales figures (right panel) for digitized and non-digitized books.[11] Before the digitization period, loans and sales moved in similar directions for digitized and non-digitized books. However, the trends changed after Google began to digitize works. Loans through the Harvard Libraries decreased for all works, but they fell more for works that were digitized. In contrast, sales of digitized works increased toward the end of the digitization period, compared to books that were not digitized.

While informative, the trends in Figure **??** do not account for the exact timing of digitization and the nature of each book. To identify the true effect of the digitization through Google Books, as well as the mechanism of this effect, we employ a more formal estimation strategy as explained in the next section.

4 Results

We follow the demand 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

$$Y_{it} = \alpha \times PostScan_{it} + \gamma_i + \mu_t + \varepsilon_{it}, \tag{4}$$

where $PostScan_{it}$ 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.

[11] We control for book fixed effects by calendar year for illustrative purposes.

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 displays the results from this specification without year fixed effects (columns 1 and 3) and with year fixed effects (columns 2 and 4). Columns 1 and 2 show that digitization through Google Books significantly decreases the number of loans through libraries. Column 1 (not including year fixed effects) suggests that making the book available on Google Books decreases Harvard library loans by about 49 ($= e^{-0.668} - 1$) percent. The impact is slightly smaller – a decrease of 38 percent – when including year fixed effects, suggesting that loans decreased over time for all books. Unlike loans, the number of *sales* is not negatively affected by digitization through Google Books. Rather, sales through traditional channels increased after digitization. When including both book and year fixed effects (column 4), we estimate an increase in sales of 0.35 ($= e^{0.297} - 1$) percent per year due to digitization. The coefficient is statistically significant at the 10% level.

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. This analysis is complicated by the fact that there is no direct control group for each digitized title because digitization happened in a staggered manner. We circumvent this problem by randomly assigning digitization years to books that weren't digitized at all, according to the distribution of digitization years in Figure **??**. Figure **??** illustrates these estimated annual impacts, lending support to the above results: digitization significantly decreases library loans at Harvard's library, whereas it seems to increase sales through other channels.

The findings in Table 2 and Figure **??** are in line with our motivating model. At Harvard's libraries, institutions for discovery are in place regardless of the presence of a digital provider, either electronically or through trained personnel. Consequently, the substitution effect dominates across all titles. In contrast, such discovery institutions are not necessarily available through other sales channels, so that the discovery effect of digitization plays an important role outside Harvard's library system.

4.3 Heterogeneous Effects

We investigate the discovery mechanism further by examining whether books of different popularities are impacted differently by the Google Books project. If search costs are otherwise high (for example for particularly obscure books), then the discovery effect of Google Books may outweigh its substitution effect. We repeat the analyses from Table 2, with an additional interaction term of the Post-Scanned variable with an indicator that is 1 if the book was checked out at Harvard's libraries more than five times in the three years before digitization, i.e. popular.[12]

Table **??** shows the results from these specifications. Consistent with expectations, the impacts of digitization vary widely for both loans and sales, as more popular books are affected more negatively. Columns 1 and 2 show 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 larger for popular works, which experience a decrease of about 81 ($= e^{-0.419-1.243} - 1$) percent, compared to a decrease of 34 percent for less popular books, according to the point estimates in column 2.

The relative impacts on *sales* are similar, although the absolute impacts are more positive. The more obscure books experience a 40 percent increase in sales – consistent with facilitated discovery outweighing substitution – whereas the estimated impact on sales of popular books is negative but not statistically significantly different from zero.

4.4 Robustness Checks

While the above results are consistent with the underlying mechanisms pointed out in previous literature and formalized in our motivating model, 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 two types of robustness checks.

We first depart from the assumption that the error terms follow a Poisson distribution, now estimating OLS regressions, with $ln(loans_{it}+1)$ and $ln(sales_{it}+1)$ as the respective dependent variables. Columns 1 and 2 of Table 3 show that the results are even stronger in this specification: digitization strongly significantly decreases library loans, and it strongly significantly increases sales.

[12] This definition identifies 1232 books as popular, accounting for 1.4% of books in the loans data, and 13.8% of books in the sales data.

Second, we match the digitized books to "similar" books – those with call numbers from the same shelf, which were not digitized purely due to their original years of publication. Using these matched samples, we repeat the analyses from Table 2 (columns 3 and 4) and from Table **??** (columns 5 and 6). Again, the results are almost identical to those in previous tables, suggesting that our original results are not driven by selection of works that were en route to becoming more or less popular.

4.5 Editions and Mechanism

While the above results suggest that digitization impacts analog demand through both discovery and substitution, additional forces are likely at play. In particular, the freely available Google Books version may allow other publishers to create and publish more and "nicer" copies of the text, allowing consumers to buy editions that did not exist previously. 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.

Table **??** shows the results of these specifications. The first two columns show 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% (column 2). Many of these editions move the title back into print in the first place, making sales possible at all.

The remaining columns of Table **??** estimate the impacts of digitization on use, controlling for the number of newly introduced editions of the work.[13] Columns 3 and 4 control for these new editions linearly, while columns 5 and 6 provide more flexible controls, including dummy variables for each number of new editions, combining observations with more than eight editions in one group.[14] Estimates from all regressions show that changes in access – while present – do not explain the previously estimated effects. Library use of the analog version decreases robustly after digitization, and some of that decrease may even be due to increased competition from other channels. Sales of printed works still increase significantly after digitization, and some of this increase can be attributed to improved access to other editions.

[13] 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.

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

5 Discussion

Some copyright holders fear that digital availability will cannibalize the use of printed works and cause financial harm. This concern has largely blocked projects that have tried to create a centralized and digital repository containing digital versions of 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. Specifically, we lay out and quantify two counteracting impacts. First, digitization may decrease demand for physical products when a work is well-known or otherwise easily discoverable – the substitution effect. Second, digitization may allow for the discovery of otherwise obscure works, thus increasing demand for those titles – the discovery effect. 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.

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. While our data do support this conclusion when discovery of new works does not depend on the digital aggregator (within the Harvard library system and – to a lesser degree – for the sales of popular titles), digitization increases sales of most other works. Combined with evidence from other studies in this literature that point to the benefits of digitization to consumers, our results suggest that existing copyright holders should be more supportive of the digitization of their catalog, especially for less popular and out-of-print works.

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. This difference means that our positive result on physical sales could be even stronger for in-copyright works, where substitution is limited to only 20 percent of the text. Our results strongly suggest that copyright holders and policy-makers should encourage the digitization and discovery of less popular in-copyright works, at least through the provision of limited amounts of text.

Finally our estimates help strengthen the value proposition of mass-digitization projects and help support their proponents such as Google Books, the Hathi Trust or the Internet Archive. While previous negotiations between these parties and copyright holders have tried to weigh the benefits to society as compared to the harm to copyright holders, our estimates suggest that this tradeoff might be relevant only when there is little

potential for additional discovery through digitization, as in the case of extremely popular titles. For a large majority of works, mass digitization projects seem to create value both for copyright holders and customers.

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. Our work also suggests multiple avenues for future work. First, we provide the first set of estimates and a research framework to evaluate the impact of digitization of books on physical sales. Future work should evaluate the robustness of our estimates in different contexts. Further, our research design looked at the impact of digitization through providing the full text of books. Future work should look at the impact of providing "snippets" and evaluate the optimal length of such limited access, balancing the positive effects from discovery and low access costs with the harmful effects to publishers from the availability of full text.

In sum, while the debate on the mass digitization of published work is complicated and involves a number of different questions, our study clarifies the important role of digitization in enabling discovery and helping copyright holders increase the sales of physical editions of their works. Our evidence therefore provides an argument for increasing access to published knowledge for all through mass digitization.

References

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

Giorcelli, M. and P. Moser (2016). Copyrights and creativity: Evidence from italian operas.

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).

Jøsevold, R. (2016). A national library for the 21st century–knowledge and cultural heritage online. *Alexandria 26*(1), 5–14.

Li, X., M. MacGarvie, and P. Moser (2018). Dead poets' propertyhow does copyright influence price? *The RAND Journal of Economics 49*(1), 181–205.

Nagaraj, A. (2018). Does copyright affect reuse? evidence from the google books digitization project. *Management Science*.

Reimers, I. (2018). Copyright and generic entry in book publishing.

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/*.

Sorensen, A. T. and S. J. Rasmussen (2004). Is any publicity good publicity? a note on the impact of book reviews. *NBER Working paper, Stanford University*.

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.

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)	7222.17	67835.59	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	802.46	8215.71	0.00	0	626610
Any Loans (0/1)	0.17	0.37	0.00	0	1
Any Sales (0/1)	0.02	0.13	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 which have 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 2003-2011). Scanned: 0/1 for books that have been scanned 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 6360 books for which sales data was collected and summary statistics are from this subgroup. Popular: 0/1 for books that have more than 5 loans before the digitization program started, i.e. between 2003-2005. Any Loans and Any Sales are 0/1 depending on whether a book was loaned or sold at least once in a given year. All sales data is generated from a sample of 6360 books for which these data is available. 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	0.113
	(0.0175)	(0.153)	(0.00167)	(0.00651)
Book FE	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes
N	792054	57240	792054	57240

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	0.113
	(0.0129)	(0.159)	(0.00167)	(0.00695)
Post Scanned x Popular	-0.303	-0.448	-0.308	0.00352
	(0.235)	(0.158)	(0.0130)	(0.0168)
Book FE	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes
N	792054	57240	792054	57240

Notes: This table presents estimates from Poisson and OLS models evaluating the impact of book digitization on loans and sales. Panel A estimates the direct effect of digitization for all books, while Panel B separately estimates effects for popular and less popular books. 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. In Panel B, Popular equals one for books that have more than 5 loans before the digitization program started, i.e. between 2003-2005. Book and year fixed effects are included in all models. Standard errors in parentheses, clustered at the book level.

Table 3. **Robustness Checks**

Panel A: Alternate Specifications

	OLS				Log Models			
	(1) Loans	(2) Loans	(3) Sales	(4) Sales	(5) Ln(Loans)	(6) Ln(Loans)	(7) Ln(Sales)	(8) Ln(Sales)
Post-Scanned	-0.0918	-0.0721	166.9	194.6	-0.0445	-0.0442	0.0600	0.0632
	(0.00325)	(0.00311)	(100.4)	(114.5)	(0.00108)	(0.00107)	(0.0177)	(0.0186)
Post X Popular		-1.681		-193.5		-0.0113		-0.0217
		(0.0559)		(114.3)		(0.0125)		(0.0412)
Book FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Popularity Effect	No	Yes	No	Yes	No	Yes	No	Yes
N	792054	792054	57240	57240	792054	792054	57240	57240

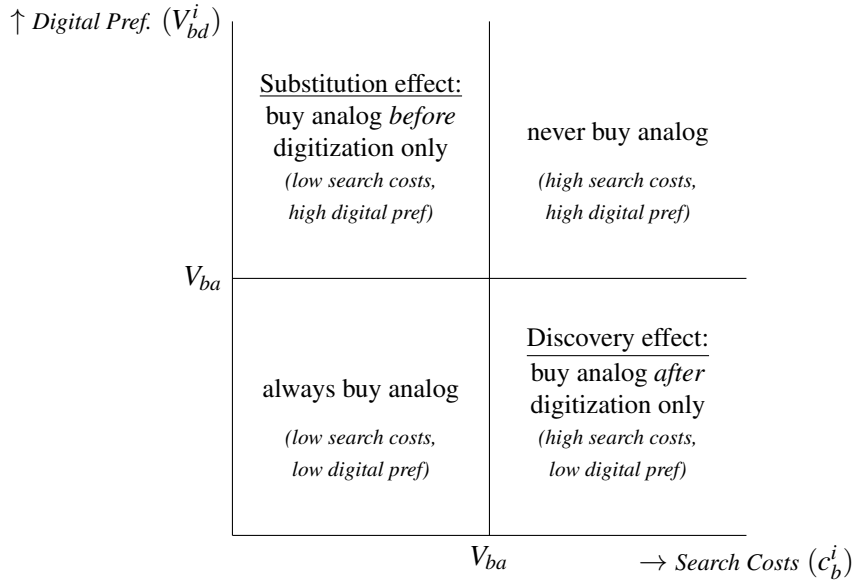
Panel B: Accounting for Editions

	Direct Effect		Accounting for Editions					
	(1) Editions	(2) Editions	(3) Loans	(4) Sales	(5) Loans	(6) Sales	(7) Any-Loans	(8) Any-Sales
Post-Scanned	2.334	0.538	-0.478	0.225	-0.473	0.299	-0.0618	0.0851
	(0.0287)	(0.0371)	(0.0150)	(0.131)	(0.0151)	(0.160)	(0.00170)	(0.00673)
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	Yes
Year FE	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	222003	222003	792054	26586	792054	26586	792054	57240

Notes: This table presents the robustness of the baseline specification. Panel A evaluates robustness to alternate specifications while Panel B investigated 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 four columns provide OLS estimates and the next four 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 (7) and (8) that rely on OLS specifications. Editions represents the number of editions available in print. Book and year fixed effects are included in all models. 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. Standard errors in parentheses, clustered at the book level.

Figures

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

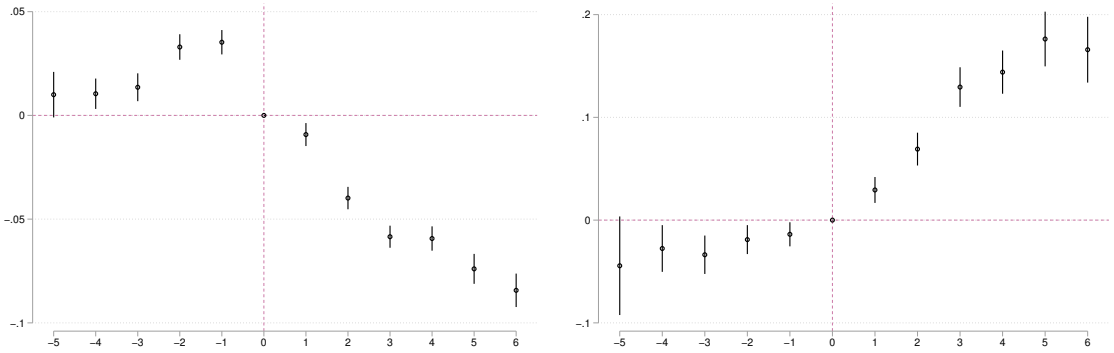


Note: This figure provides an illustration of predictions from the theoretical framework. The framework models an individual customer i's decision to purchase an analog version of the book (a physical copy) as a function of their search costs c_b^i (x-axis) and preference for digital copies V_{bd}^i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

Figure 2. **Visualizing the Effects of Digitization on Scanned Books**

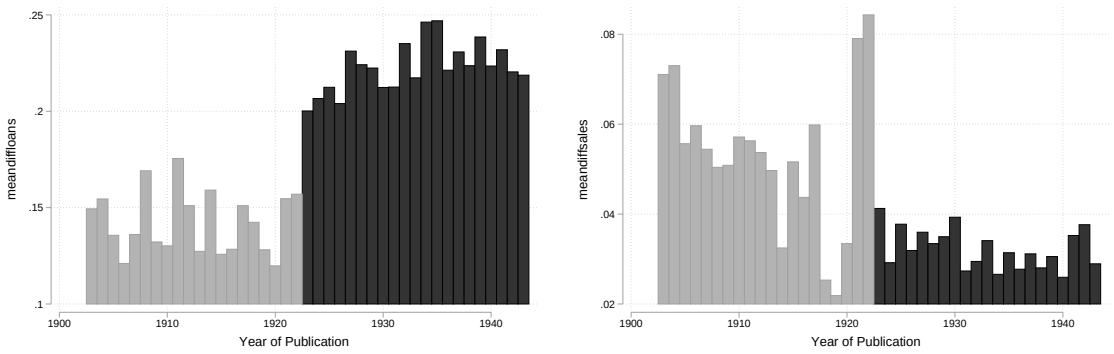
Panel A: by Timing of Book Digitization

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



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

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



Note: This figure provides visual illustrations of the main results. In Panel A, we plot time-varying estimates drawing on the baseline difference-in-difference specification in Table 2, Panel A (OLS). Specifically, we estimate a series of co-efficients according to an event study specification where the main binary independent variable is replaced by a series of leads and lag indicators which represent number of years before or after a particular book is scanned. The main dependent variables here are Any-Loans (i) or Any-Sales (ii).

In Panel B, we explore the impact of the digitization program on a cross-sectional sample of a subset of 57,316 english language books of which only those published before 1923 are digitized. Books published after 1923 were not digitized because they were likely protected by copyright and so Harvard could not legally allow Google Books to digitize them. 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 average likelihood that books in a certain publication year 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.