Nagaraj and Reimers (2021): Replication and Extension

David Scolari, John Bowman, Blake Lin

May 2022

Abstract

1 Introduction

In their 2021 study, Abhishek Nagaraj and Imke Reimers examine whether digitization of books may actually increase, not decrease, sales of their physical versions when this digitization is accompanied by full-text search technology. They find an extensive literature that establishes the tendency of free or low-cost provisions of media "cannibalizing" sales of older formats. However, prior to Nagaraj and Reimers 2021, there has been little work investigating the potential for digitization of antiquated media to improve search functions and hence increase sales for physical versions.

Part of their empirical approach uses loan activity for books within the Harvard library system. They find that, at least within the Harvard libraries, the effect of book digitization is negative on loans for those books that were digitized. Nagaraj and Reimers explain that, within a university library, a negative result makes sense because any search cost decrease due to digitization will be muted by the already robust search services that the library provides.

We aim to replicate Nagaraj and Reimers's identification strategy for the effect of digitization within the Harvard Library system, which relies on two-way fixed effects (TWFE). We will then apply recently developed difference in differences (DID) techniques which address the bias introduced by TWFE, specifically Goodman-Bacon's decomposition of the DID estimator under differential treatment timing and Calloway and Sant'Anna's DID estimator.

1.1 Still need to address:

Such an effect could be particu- larly salient in the market for books, where digital distribution can be accompanied by technologies that allow consumers to search through the full-text and discover (and buy) new content that would be otherwise hard to find (Ellison and Ellison, 2018).

the Google Books digitization project. Launched in 2005, Google Books is one of the landmark projects of the digital age, with commentators likening it to a "modern-day Library of Alexandria" (Somers, 2017). Google Books did not just scan a book's textual material but also made it searchable via optical character recognition (OCR) technology through its "Google Book Search" feature (referenced by Cambridge University Press in the epigraph). Further, a large portion of the Google Books corpus included less well-known and older books (including public domain content) that are of significant consumer interest but have become forgotten over time.

We tackle the empirical challenges through a unique natural experiment leveraging a research partnership with Harvard's Widener Library, which provided books to seed the Google Books program. The digitization effort at Harvard only included out of copyright works, which – unlike in-copyright works – were made available to consumers in their entirety. This allows us to fairly assess the tradeoff between cannibalization (by a close substitute) and discovery (through search technology). Owing to the size of the collection, book digitization (and subsequent distribution) at Widener took over five years, providing signifi- cant variation in the timing of book digitization.

[EXPLAINING HOW THE MUTED EFFECT HELPS THEM IDENTIFY THE OVERALL EFFECT] Next, we examine the importance of the discovery and substitution channels by studying two paral- lel settings where one of these two effects is muted. First, we investigate the effects of digitization on loans within the Harvard system. In this setting, the discovery effect is muted since Harvard students and professors already had access to alternate discovery mechanisms through library services. In this setting, digitization reduces rather than increases demand, measured as loans within Harvard.

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

1.2 How Nagaraj and Reimers think digitization is working wrt demand (and supply) for books

How might digital distribution increase the sales of physical books? For end customers, the question of whether the digitization of books increases or reduces demand for physical works depends on two counter- acting forces. The first is the substitution effect of digital distribution as a competitor for existing, physical products as studied in the literature. Some consumers who would otherwise consume physical copies will switch to digital versions, driving the substitution effect. This is likely to happen when a consumer's search costs are low, and when she has a taste for digital consumption, and it may be less relevant when books are

However, Google Books might stimulate a discovery effect due to increased awareness and searchability (see appendix Figure D.1.) Specifically, some consumers may start consuming the physical version for the first time, after digitization lowers search costs for books. This is likely to happen if they were made aware of a book through Google Books' search engine and prefer to purchase physical copies rather than read online. This second mass of consumers will drive the discovery effect. The net effect of digitization is ambiguous and depends on the magnitude of these two margins.

The tradeoff between substitution and discovery further differs for different margins of books and con- sumers. Notably, for popular books, already well-known to consumers (e.g. The Wealth of Nations), the substitution effect is likely to dominate. On the other hand, obscure books are

likely to benefit from discov- ery, and unlikely to face the costs of substitution. The effect of Google Books on demand should therefore be more positive for less popular books. In addition, if consumers discover a particular author through a dig- itized copy, they might also seek out other books by the same author, even if these have not been digitized. Therefore, digitization might lead to an increase in physical sales for non-digitized works of a digitized author as well.

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

To summarize, our theoretical predictions are threefold. First, digital distribution allows consumers to search for topics they are interested in and discover works previously unknown to them. Second, if the digital medium offers a poor substitute for a physical book, demand for physical works is more likely to increase. Finally, digital provision will also allow publishers (especially small and independent ones) to identify new material and introduce new editions. Our theoretical arguments speak to past work on the impact of digital distribution on demand and supply for physical products. On the demand side, our arguments about the role of digital distribution in enhanc- ing consumer discovery add to work on the effects of digital technology on the diversity of consumption patterns (Brynjolfsson et al., 2006; Kumar et al., 2014; Holtz et al., 2020). This work shows that digital distribution channels tend to change consumption patterns by helping consumers discover more novel and niche products. Related work has looked at the complementary effects of news content sampled via social networks, which drives traffic to news websites by helping consumers discover specific news articles (Chiou and Tucker, 2017; Sismeiro and Mahmood, 2018). However, this literature focuses on how digital channels change consumption patterns and has not explored how online access coupled with access to digital search and discovery tools affects demand for and supply of the same product in physical form. Our theoretical and empirical analyses extend this work in this direction.

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

2 Literature Review

The motivation for this project comes from literature which examines the effect of free or low cost digital distribution of information goods on the demand on their physical counterparts. (Smith and Zentners 2016 review(1).

The is a breadth of resources which investigate the impact of illegal downloads. The bulk of this research focus on the impact of file sharing on music sales (2), while others look at the impact on the film industry(3). Recently, researchers has incorporated legal forms of cheap digital media distribution such as online streaming and found that it also shifted demand of more expensive substitutes(4). In these industries, the literature overall tends not to find that digital distribution shifts supply for the physical counterpart, which is perhaps due to digital distribution doe not explicitly improve the information environment.

There are several key differences to low cost digital distribution for print media. Digitized print media that is scanned in by the entire text can be searched by the entire text, which enables a more complete match between the content and the users query. This match quality can increase demand because of the higher match quality.(5) Specifically, readers are able to find books that they would not want to read or not even know existed. Consumers who want a more user friendly format besides the low quality reader buy the physical book, entering the market.

Another method of acquisition is checking the book out form the library. Books that may have been left on the shelf are now being checked out and read because people were made aware of their existence. However scanning books could also have the opposite effect on the demand for them. A book that's easily available on google books could save a potential consumers a trip the book store or library, or the wait for the online order to deliver. Depending on what type of book the consumer seeks, it could even be more advantageous to have it in a digital format. Large textbooks for example offer a much more utility if they book itself can be queried for specific information, not to mention forgone cost of the physical version.

Since the theoretical argument remains ambiguous, an empirical study into the causal effect is warranted. When designing a causal research question, model specification is key. The most common specification is Two Way Fixed Effects, which compares outcomes for observations with the treatment to observations without the treatment. It is a simple model design, which is flexible to various controls, but has a critical flaw. In circumstances when there are multiple treatment groups, Two Way Fixed Effects is specified in a way which causes observations which have been treated to be compared to newly treated observations. This contaminates the leading and lagging indicators when other assumptions are imposed like parallel trends and limited anticipation of the treatment.(6) The remedy for this issue is found in Callaway Sant'anna (7) specification, which bins the treated observations into time treated cohorts. This separates out the treatment effects and prevents overlapping. It also enables us to check the plausibility of prior trends and see how the treatment effect changes over time.

3 Data

The raw data we use to replicate Nagaraj and Reimers 2021 is the loan activity for books within the Harvard library system between 2003 and 2011. There were a total of 88,006 books loaned during that time period. We build our panel using each of these 88,006 unique books as our panel id and counting the number of loans it experiences between 2003 and 2011. This panel allows us to directly replicate part Nagaraj and Reimers 2021.

We also observe a borrower identifier in the raw data, allowing us to count loan events for different types of borrowers, such as Harvard faculty and students. We use these borrower IDs to make panels that, instead of counting loans made by all types of borrowers, only count loans made by specified borrower types. For these panels, we still use the 88,006 unique book IDs as our panel units, so the datasets do not change in number of observations, just in their method for counting

loan events. This data that includes borrow type specified loan counts is used for our extension to Nagaraj and Reimers 2021 and does not appear in their paper at all.

3.1 Still need to address:

We tackle the empirical challenges through a unique natural experiment leveraging a research partnership with Harvard's Widener Library, which provided books to seed the Google Books program. The digitization effort at Harvard only included out of copyright works, which – unlike in-copyright works – were made available to consumers in their entirety. This allows us to fairly assess the tradeoff between cannibalization (by a close substitute) and discovery (through search technology). Owing to the size of the collection, book digitization (and subsequent distribution) at Widener took over five years, providing signifi- cant variation in the timing of book digitization.

4 Methodology

5 Results

In first and third columns of Table 1, we directly replicate a result published in Nagaraj and Reimers. Using both log-inflated loans (log-OLS) as well as a binary indicating whether or not a book was loaned in a given period (LPM) as outcome variables, as well as book and year-location fixed effects, Nagaraj and Reimers find a negative and statistically significant effect of digitization for books within the Harvard Library system. This results suggests that digitization causes a substitution effect that is stronger than the demand increase it induces by reducing search costs. We expect the substitution effect to be stronger for this set of books because, within the Harvard library system, there are already robust services to help borrowers find books, such as librarians and online databases. Nagaraj and Reimers propose that the search cost effect of digitization is "muted" within the Harvard library system because search costs are already low citenr2021.

Table 1: Nagaraj and Reimers Replication

	(1)	(2)	(3)	(4)
	log-OLS	log-OLS	LPM	LPM
Post-Scanned	-0.0511***	-0.0518***	-0.0613***	-0.0627***
	(0.00152)	(0.00146)	(0.00170)	(0.00163)
Book FE	Yes	Yes	Yes	Yes
Year FE	No	Yes	No	Yes
Year-Location FE	Yes	No	Yes	No
N	792054	792054	792054	792054

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

The second and fourth columns display the results of similar models that use year fixed effects instead of year-location fixed effects. These models are part of our extension and not found in Nagaraj and Reimers. We simplify the time fixed effects in order to better align Nagaraj and Reimers's DID specification with Goodman-Bacon's decomposition and Calloway and Sant'Anna's

estimator. Since the aim of this extension is to identify the effect of digitization using these new methods, we need a base two-way fixed effects specification that can comply with them. We present these simplified model estimates along side Nagaraj and Reimer's original results to show that dropping the location element of the time fixed effects has only a small impact on the estimates' significance, sign, and magnitude.

Table 2: Additional Results

	(1) log-OLS	(2) log-OLS	(3) log-OLS
Post-Scanned	-0.0518***	-0.0669***	-0.0655***
	(0.00146)	(0.00132)	(0.00131)
Faculty × Scanned		0.0224***	
		(0.00231)	
Doctorate × Scanned			-0.0406***
			(0.00303)
Masters × Scanned			0.00198
			(0.00338)
Undergrad × Scanned			0.0148***
			(0.00248)
Book FE	Yes	No	No
Book-Borrower FE	No	Yes	Yes
Year FE	Yes	Yes	Yes
N	792054	1584108	3168216

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

6 Conclusion

Table 3: Bacon Decompositison

2x2 Type	Avg. Estimate	Weight
All Borrowers		
Earlier vs Later Treated	-0.195	0.341
Later vs Earlier Treated	-0.184	0.250
Treated vs Untreated	-0.590	0.409
Doctoral Students		
Earlier vs Later Treated	-0.084	0.341
Later vs Earlier Treated	-0.468	0.250
Treated vs Untreated	-0.693	0.409
Faculty		
Earlier vs Later Treated	-0.213	0.341
Later vs Earlier Treated	-0.184	0.250
Treated vs Untreated	-0.485	0.409
In-Building		
Earlier vs Later Treated	-0.213	0.341
Later vs Earlier Treated	-0.184	0.250
Treated vs Untreated	-0.485	0.409
Masters Students		
Earlier vs Later Treated	-0.179	0.341
Later vs Earlier Treated	0.047	0.250
Treated vs Untreated	-0.371	0.409
Undergraduate Students		
Earlier vs Later Treated	-0.048	0.341
Later vs Earlier Treated	0.026	0.250
Treated vs Untreated	-0.242	0.409

Table 4: TWFE vs. Group CS Estimators

Borrower Group	TWFE	Calloway Sant'Anna	Significance	Sign Change
All Bowrrowers	-0.0518	-0.0388	More	No
	(0.00728)	(0.000342)		
Doctoral Students	-0.107	-0.025	More	No
	(0.0169)	(0.00305)		
Faculty	-0.0515	-0.0267	More	No
	(0.0126)	(0.000869)		
In-Building	-0.0515	-0.0267	More	No
	(0.0126)	(0.00196)		
Masters Students	-0.066	-0.0244	More	No
	(0.0149)	(0.00394)		
Undergraduate Students	-0.05	-0.0209	More	No
	(0.0116)	(0.0039)		

Standard errors in parenthesis