Nagaraj and Reimers (2021): Replication and Extension

David Scolari, John Bowman, Blake Lin

May 2022

Abstract

1 Introduction

The Google Books project, launched in 2005, is one of the landmark projects of the digital age, not only scanning the text of a gargantuan corpus of books, but also making that textual content searchable by consumers. 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.

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

Table 1: Treatment Cohorts

	Cohort	Number of Books
1	Never Treated	50,289
2	2005	5,746
3	2006	7,449
4	2007	8,769
5	2008	13,207
6	2009	2,546

5 Results

5.1 Replicating Nagaraj and Reimers

In the 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.

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

Table 2: Nagaraj and Reimers Replication

	(1) log-OLS	(2) log-OLS	(3) LPM	(4) 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

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

dropping the location element of the time fixed effects has only a small impact on the estimates' significance, sign, and magnitude.

5.2 Borrower Effects

For the baseline TWFE (column 1), the Average Treatment Effect on the Treated (ATT) of a book getting scanned into google books on log inflated loans was -.0518 and significant at the 99% level. Put differently, the causal effect of scanning a book on checkouts was 5% fewer checkouts from Harvard libraries. This ATT from the overall population shows that books available on google books act as substitutes for their physical counterparts. It could indicate either that the superior convenience and/or cheaper cost of low-cost digital information goods exceeds the value lost from the unwieldy digital format.

When the population is segmented by faculty group membership, the post-scanned causal effect becomes more pronounced while retaining its significance at the 99% level. This is due to the extraction of the positive ATT for scanned books checked out by faculty. The ATT implies that faculty members use google books more as an information source for later physical acquisition. From this, it follows that the negative checkout rates are driven by non-faculty members who are using google books as a substitute.

This effect is further decomposed among the student groups (column 3) while retaining their 95% levels of significance. Among masters students and undergrads the usage of google books is similar to that of faculty members. For these populations which have positive ATTs, google books is used as a resource finder as opposed to a format for using those resources. Doctoral students on the other hand, exhibit behaviors which generally use google books as a format for consuming those information resources rather than a research tool. Why doctoral students diverge from their other student counterparts is not clear. One possible narrative to explain why doctoral students exhibit this behavior is a preference held by the group at large to use the cheap digital format and forgoes the opportunity cost of going to the library where the physical resource resides. This tracks with the increased responsibilities usually associated with doctoral studies compared to other education levels, such as Research Assistance, teaching undergraduate courses, tutoring, etc., which could lead to a more acute scarcity of time.

Perhaps the most interesting ATT for the third model is the base post scanned causal effect. Here the magnitude of the baseline ATT a closer to the magnitude in the faculty/non-faculty than

the non-grouped models. Since the final model groups students together, faculty members are classified as non-students. We learned in the second model that faculty members use google books as a reference resource, yet even with that effect in the base post scanned ATT, there is still a significant negative ATT. From this we can infer that for non-faculty non-students are the main drivers of google books as a platform for viewing digital resources substituting away from physical mediums. The reason for this is up for speculation. One possible story is that many of these non-student/non-faculty library patrons face increased barriers to receive credentials to use these libraries, leading to reduced checkout rates for these groups. If getting a Harvard library card is more difficult for non-student/non-faculty members, it makes sense for them to switch to google books if the resource they need is available on the platform, forgoing an increased opportunity cost compared with students or faculty members.

Table 3: 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

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

When the population is segmented by faculty group membership, the post-scanned causal effect becomes more pronounced while retaining its significance at the 99% level. This is due to the extraction of the positive ATT for scanned books checked out by faculty. The ATT implies that faculty members use google books more as an information source for later physical acquisition. From this, it follows that the negative checkout rates are driven by non-faculty members who are using google books as a substitute.

This effect is further decomposed among the student groups (column 3) while retaining their 95% levels of significance. Among masters students and undergrads the usage of google books is similar to that of faculty members. For these populations which have positive ATTs, google books is used as a resource finder as opposed to a format for using those resources. Doctoral students on

the other hand, exhibit behaviors which generally use google books as a format for consuming those information resources rather than a research tool. Why doctoral students diverge from their other student counterparts is not clear. One possible narrative to explain why doctoral students exhibit this behavior is a preference held by the group at large to use the cheap digital format and forgoes the opportunity cost of going to the library where the physical resource resides. This tracks with the increased responsibilities usually associated with doctoral studies compared to other education levels, such as Research Assistance, teaching undergraduate courses, tutoring, etc., which could lead to a more acute scarcity of time.

Perhaps the most interesting ATT for the third model is the base post scanned causal effect. Here the magnitude of the baseline ATT a closer to the magnitude in the faculty/non-faculty than the non-grouped models. Since the final model groups students together, faculty members are classified as non-students. We learned in the second model that faculty members use google books as a reference resource, yet even with that effect in the base post scanned ATT, there is still a significant negative ATT. From this we can infer that for non-faculty non-students are the main drivers of google books as a platform for viewing digital resources substituting away from physical mediums. The reason for this is up for speculation. One possible story is that many of these non-student/non-faculty library patrons face increased barriers to receive credentials to use these libraries, leading to reduced checkout rates for these groups. If getting a Harvard library card is more difficult for non-student/non-faculty members, it makes sense for them to switch to google books if the resource they need is available on the platform, forgoing an increased opportunity cost compared with students or faculty members.

6 Conclusion

Table 4: Bacon Decompositison

2x2 Type	Avg. Estimate	Weight
All Borrowers		
Earlier vs Later Treated	-0.051	0.341
Later vs Earlier Treated	-0.016	0.250
Treated vs Untreated	-0.090	0.409
Doctoral Students		
Earlier vs Later Treated	-0.008	0.341
Later vs Earlier Treated	-0.024	0.250
Treated vs Untreated	-0.148	0.409
Faculty		
Earlier vs Later Treated	-0.031	0.341
Later vs Earlier Treated	-0.010	0.250
Treated vs Untreated	-0.076	0.409
In-Building		
Earlier vs Later Treated	-0.031	0.341
Later vs Earlier Treated	-0.010	0.250
Treated vs Untreated	-0.076	0.409
Masters Students		
Earlier vs Later Treated	-0.017	0.341
Later vs Earlier Treated	-0.005	0.250
Treated vs Untreated	-0.093	0.409
Undergraduate Students		
Earlier vs Later Treated	-0.013	0.341
Later vs Earlier Treated	0.001	0.250
Treated vs Untreated	-0.069	0.409

Table 5: 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