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 to "cannibalize" 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. They suggest two possible explanations for how book digitization might effect the sales of physical books. The first is through a substitution effect, where a book's scanned text offer consumers a substitute to the physical version. The second is by aiding discovery and thereby increasing demand for a book, physical or otherwise. For instance, if a consumer prefers to read physical copies of books but is made aware of a book through Google Books, then they might go and purchase (or borrow) the physical copy. Since these two effects oppose each other, the net effect of digitization on physical book sales is ambiguous and must be addressed empirically.

Part of Nagaraj and Reimers's empirical approach uses loan activity for books within the Harvard library system. These books were all out of copyright, which meant that when they were digitized by Google, they were made available to consumers in full, providing a comparable alternative to their hardcopy counterparts. The full digitization of this set of books at Harvard took over five years, providing providing a natural experiment where each book's loan activity is observed both pre and post digitization. Nagaraj and Reimers find that, at least within the Harvard libraries, the effect of book digitization on loans is negative for those books that were digitized. They 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. Since the digitization took five years, their identification relies on two-way fixed effects (TWFE) to estimate the difference in differences (DID) with variation in treatment timing. We will then apply recently developed 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.

This paper will proceed with a literature review outlining both the literature that Nagaraj and Reimers contribute to as well as literature pertaining to methods that address bias in TWFE. We will then give a brief overview of the loan activity data and the panels we create from it. Next, we will walk through the TWFE specification, the bias it introduces into it's DID estimator, and how new methods aim to address that bias. Lastly, before concluding, we present the results of our replication and extension to Nagaraj and Reimers.

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

There are several key differences to low cost digital distribution for print media. Digitized print media can be searched by the entire text, which enables a heightened compatability between the content and the users query. This enhanced match quality can increase demand for the physical resource.(5) Specifically, readers are able to find books that they would not want to read or not even know existed, then aquired physical copies. These Consumers who want a more user friendly format than 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 thanks to the reaserch utility digital formats. 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 more utility if 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 infrence 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 the Callaway Sant'anna (7) specification, which bins the treated observations into time treated cohorts. This separates out the treatment effects and prevents the overlapping observed in the Two Way Fixed Effects.

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. The full digitization of this set of books at Harvard took over five years, providing providing a natural experiment where each book's loan activity is observed both pre and post digitization. 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.

4 Methodology

By estimating the effect of digitization on book loans in the Harvard library system, Nagaraj and Reimers attempt to determine how the demand for books is effected by digitization in a setting where the search cost benefits of digitization are muted by the already robust tools for locating books that exist within a university library. They expect to find negative results that indicate the substitution effect of digital books is dominating the discovery effect.

Books are scanned by Google over a five year period. Most of our 88,006 books are never scanned, and between about two and thirteen thousand books are scanned each year. Because of this digitization rollout, our treatment is assigned with differential in timing, meaning that a book might be part of the untreated group in one year and part of the treated group in another year. This treatment rollout is summarized in Table[].

Nagaraj and Reimers use a modified TWFE specification to identify their DID estimator. Generally, TWFE estimates the difference in differences estimator with a regression of the following form

$$y_{it} = \beta_0 + \delta D_{it} + \beta_1 X_{it} + \alpha_t + \alpha_i + u_{it} \tag{1}$$

where D_{it} is a dummy that indicates whether unit i is treated during period t. Nagaraj and Reimers deviate from this form slightly in their paper by using the grouped variable of year and shelf location as their time variable.

While TWFE has for years been the default method for estimating DID with differential timing, recent work in econometrics cite()()() has identified bias that TWFE induces on its estimate. This bias is best show by the Bacon decomposition, which is shown in table [].

Two-way Fixed Effects (TWFE) is the standard Difference-in-Differences estimation method. However, its efficacy is limited to situations when treatment occurs simultaneously. Since our data set contains multiple time periods and variations in treatment timing - the books were scanned in different years, here we would try using Bacon decomposition to solve the problem of using late-treated units compared to early-treated units. By using the binary treatment variable, we will

Table 1: 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

re-estimate the effect of scanned books on borrowing rate by coding a book as “treated” if at any time of that year it had been scanned into digital form.

For simplicity of modeling and interpretation, scanned will be the only variable on the right-hand side, we would not include other covariates. Second, we aggregate the panel units by treatment cohort for computational purposes. To see the treatment effect on different groups of borrowers, we filtered the data into different datasets based on borrower types.

The table below is showing the weight and estimates of each type of borrower:

Looking at the table from the “All borrowers” result, we can see in the Bacon decomposition that less than half of the TWFE parameter estimate is coming from comparing the treatment books to a group of never-treated books. The average DD estimate for that group is -0.59 with a weight of 0.409, so the influence of later to early treated in the mix is rather large and decreases the treatment effect in the TWFE estimate. The influence of later to early treated remains the same for other types of borrowers, that the estimate of treated vs. never-treated being the highest but got pulled down because of differential timing groups.

When we have this differential timing in our treatment assignment, we need to alter our interpretation of the DID estimator. Instead of a single 2x2 DID, we take the weighted average of multiple 2x2 DIDs to compute our estimator. This begs the question: How exactly does TWFE weight the individual 2x2 estimators?

The weighting of the 2x2 DIDs is the subject of Andrew Goodman-Bacon's 2019 paper. He shows how the weights that OLS imposes on the 2x2 DIDs introduces bias into the TWFE estimator. Calloway and Santa'Anna 2021 proposes an alternative weighting that addresses this bias. We will discuss the Bacon decomposition of the TWFE in the upcoming subsection, followed by a comparison between a TWFE specification of Lott and Mustard's analysis and Calloway and Santa'Anna's DID (CS) specification.

The CS estimator splits treatment units into cohorts by assignment date. It then finds the ATT for each cohort g over the rollout period. It takes a weighted overage of these cohort ATTs to give an overall ATT estimate.

Table 2: Treatment Cohorts

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

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 3: 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 4: 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.

5.3 extension to new estimators

Now, we would try using the Callaway-Sant'anna (CS) method to avoid the problem of late-to-early comparison by only using the never or not-yet treated as controls group through subsetting dataset. This method would split the observations with the same time of treatment into cohorts, then estimate the treatment effect based on the cohorts. The following table compares the TWFE model to CS in capturing the effect of the scanned treatment on borrowing rate loginflloans. The first row shows the overall "scanned effect" for all borrowers, and the rest is showing the treatment effect of other types of borrowers using different sets of observations.

We can see that in all of the groups, TWFE overestimated the treatment effect. By using the CS estimation, the sign of the treatment effect maintains the same but the significance of the estimator increases in all of our groups. The reduction of treatment effect by using the CS estimation range from 1.3% to 8.2%, the lowest at 1.3% being the effect for all borrowers and 8.2% for doctoral students.

6 Conclusion

Expansions on the two way fixed effect models can be achieved through additional data on group membership. We found that nonstudent/nonfaculty library patrons were primarily responsible for the reduced checkout rates for treated books. One alternative angle is decomposing that groups to see if it is driven by some subgroup. This will be constrained by what information is collected by the library on these patrons as well as what they are allowed to share. This study could also be applied to non university libraries, which will have different patron population compositions. If

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

non student/non faculty ATT remains consistent, this could cause even more drastically reduced checkout rates, since public libraries serve the general public rather than academic populations.

Another alternate approach is to compare checkoput rates at other universites. Are the group spesific trends consistent between university populations, or are negative ATTs for doctoral checkoput rates a unique quality of Harvards doctoral students? Do faculty members use google books as a refrencing resource in other universities, or is the technology not used in this way. Faculty age could be taken into account as well. Perhaps there is a decreased ATT due to google books being a relativly new refrencing tool which is employeed at different rates for different age demographics.

Being a platform, google books has a network effect. A network effect is the dynamic in which the utility of a good increases as the use of that good increases. In this scenario, as more books are scanned, more people will defer to google books as the information resource, forgoing other options like libraries or book stores. It may be worth exploring if the total number of books scanned into google books depresses checkout rates/purchase rates for non treated books. This could have a negative effect on the information environment as it would cause non scanned books with greater compatiblity to the querying user going undiscovered in favor of less compatible scanned books.