# Replicating and Extending Nagaraj and Reimers, 2021: Using New DID Techniques to Estimate the Effects of Book Digitization on Library Loans

David Scolari, John Bowman, Blake Lin

May 2022

**Abstract**

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. We aim to replicate and extend Nagaraj and Reimers's identification strategy by applying recently developed DID techniques which address the bias introduced by TWFE. Specifically, we implement Goodman-Bacon's decomposition of the DID estimator under differential treatment timing and Callaway and Sant'Anna's DID estimator. We also explore ATTs for different borrower groups. Our extension serves mostly to confirm the validity of the original researchers' results, as the CS estimator does not flip the sign of any of the estimates produced by TWFE. Moreover, effects are similar across the borrower groups we explored at Harvard.

## 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 (Nagaraj & Reimers, 2021). 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.

Nagaraj and Reimers 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, even if a consumer prefers to read physical copies of books, they might be made aware of a book through Google Books and 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 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 (Nagaraj & Reimers, 2021). 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 Callaway 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. Before concluding, we present the results of our replication and extension to Nagaraj and Reimers.

## 2 Literature Review

The motivation for Nagaraj and Reimers's 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 & Zentner, 2016). The breadth of resources on this topic investigate the impact of illegal downloads. The bulk of that research covers file sharing on music sales (Bounie, Bourreau, & Waelbroeck, 2006), while other studies look at the impact on the film industry (Rob & Waldfogel, 2007). Recently, researchers have incorporated legal forms of cheap digital media distribution such as online streaming and found that it negatively shifted demand of more expensive substitutes (Aguiar & Waldfogel, 2018). In these industries, the literature overall tends to find that digital distribution negatively shifts demand for the physical counterpart, which is perhaps because digital distribution does not improve the information environment.

However, when there is a strong improvement to the information environment, there is reason to believe that digitized media might increase demand for physical media. This is the effect that Nagaraj and Reimers aim to test in their 2021 paper. They employ several causal inference technique to do so, and we intend to expand on that work in this paper.

When designing a causal inference 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 (Sun & Abraham, 2021).The remedy for this issue is found in the Callaway Sant'anna (Callaway & Sant'Anna, 2021) 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 count 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 between 2005 and 2009. 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 1.

Table 1: 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 |

Nagaraj and Reimers use a modified TWFE specification to identify the ATT. 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 the dummy variable, $D_{it}$, indicates whether unit $i$ is treated during period $t$ and $\delta$ is the DID estimator. Nagaraj and Reimers deviate from this form slightly in their paper by using the grouping of year and shelf location as their time variable. In their regression, their outcome, $y_{it}$, is log-inflated loans and $D_{it}$ indicates whether a book has been scanned into Google Books in a previous time period. They do not include any covariate controls.

In our extension to Nagaraj and Reimers, we both simplify and expand upon their regression. We simplify by using year fixed effects instead of year-location fixed effects. The purpose of this is to allow our TWFE regression to work with a Bacon decomposition and with the Callaway/Sant'Anna estimator. We expand upon their regression by adding borrower type dummies to the model.

$$y_{it} = \beta_0 + \gamma_0 Borrower + \delta D_{it} + \gamma_1 Borrower \times D_{it} + \beta_1 X_{it} + \alpha_t + \alpha_i + u_{it} \tag{2}$$

Here, $\delta + \gamma_1$ identifies the ATT with respect to loans made by a particular borrower group, such as faculty members and different types of students at Harvard.

Our extension also aims to use updated DID techniques in lieu of TWFE. While TWFE has for years been the default method for estimating DID with differential timing, Goodman-Bacon 2021 shows that this method introduces bias into its estimate (Goodman-Bacon, 2021). The bias comes from the fact that the weights that OLS imposes on the individual 2x2 DIDs favor periods where a large number of units have the treatment assigned as well as 2x2s between units whose proportion of time treated over the span of the panel is close to 0.5. We can see this bias in the "Bacon decomposition". We show this bias in Nagaraj and Reimer's regression with a Bacon decomposition in the results section.

We also estimate the Callaway/Sant'Anna ATT of book digitization within the Harvard library system. This estimator uses the following equation to identify the ATT.

$$ATT_{(g,t)} = E\left[\left(\frac{G_g}{E\left[G_g\right]} - \frac{\frac{\hat{p}(X)C}{1-\hat{p}(X)}}{E\left[\frac{\hat{p}(X)C}{1-\hat{p}(X)}\right]}\right)(Y_t - Y_{g-1})\right]$$

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. This method of estimating DID addresses the bias caused by TWFE (Callaway & Sant'Anna, 2021).

# 5 Results

## 5.1 Replicating Nagaraj and Reimers

In the first and third columns of Table 2, 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 (Nagaraj & Reimers, 2021).

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

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

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 Callaway 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.

## 5.2 Borrower Effects

Table 3 shows model estimates for ATTs by borrower type. 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 is the same result shown in Table 2, column 2. 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.

The next two columns break down the effect of digitization by borrower type. Unsurprisingly, we see that all of the effects are negative, indicating that both faculty and students alike decrease their loan activity when books are digitized. We can interpret this to mean that the substitution effect dominates any discovery effect taking place for these groups. Indeed, it makes sense that the discovery effect is muted for students and faculty at Harvard, as they are likely to be highly adept

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

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

at using the library's book finding services.

Moreover, although not to be read too deeply into without significance tests, there appears to be a ranking of effects within students that aligns with reality. Doctoral students have the most negative ATT perhaps because they have the most skill when it comes to using the library search functions compared to the other sets of students. Because of this, Google Books does not lower their search costs very much at all, so their ATT is heavily dominated by the substitution effect. Masters and undergraduate students have similar ATTs, likely not significantly different from one another. But to offer a narrative to explain this slight difference, perhaps undergrads have a more negative ATT because they are on campus for more years than masters students, allowing them to better familiarize themselves with the library's search functions.

## 5.3 New Estimators

As stated in our methodology section, we aim to re-specify Nagaraj and Reimers TWFE model with new DID techniques. Table 4 summarizes the Bacon decomposition for a TWFE model that uses loan activity for different borrower types as the outcome and scanning into Google Books as the treatment. The rows labeled "All Borrowers" decompose an estimator similar to Nagaraj and Reimer's original result (they only differ from NR by swapping year-location FEs for year FEs). Due to our panel size exceeding our computing constraints, we aggregate the panel units by treatment cohort. The Bacon decomposition shows us the weights imposed on the individual 2x2s by OLS.

We see that Earlier vs Later 2x2s are weighted more heavily than Later vs Earlier 2x2s. We also

see that Treated vs. Untreated 2x2s have the heaviest weight. There is no economic justification for this weighting. We also see that average estimates are different (middle column). For this particular set of TWFE regressions, the weights may not be a big problem since the average estimates are, with a few exceptions, the same sign and of similar magnitude.

Now, we use the Callaway/Sant'Anna (CS) method to address the spurious weights imposed on our estimates by TWFE. The following table compares the TWFE model to CS in capturing the effect of the scanned treatment on borrowing rate log inflated loans. The first row shows the overall "scanned effect" for all borrowers, and the rest show the treatment effect of other types of borrowers using different sets of observations. The TWFE ATTs estimated here differ from those shown in Table 3 in that borrowers are separated by filtering the data and running separate regressions on the filtered panels, where the models in Table 3 use dummy variables to estimate ATTs for different borrowers.

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 |

We can see that in all of the groups, TWFE estimates ATTs of higher magnitude than CS. By using the CS estimation, the sign of the treatment effect remains the same but the significance of the estimator increases in all of our groups. The reduction of treatment effect by using the CS

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

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.

Perhaps the most significant finding is that the CS estimator severely reigns in the ATT for doctoral students, which was estimated to be about twice that of the other borrower types by previous estimation methods. This finding is important because TWFE estimated the doctoral ATT to be somewhat of an outlier among the other borrower groups, however, CS estimates it to be commensurate compared to the other groups. Looking back at the Bacon Decomposition in Table 4, we might expect the CS estimator to have this effect because TWFE heavily weights the Treated vs Untreated DID estimates, which are very high for doctoral students.

Overall, the CS estimator does not change the story the Nagaraj and Reimers tell with their model. For all borrower types of interest, the CS estimator decreases the magnitude of the ATT of digitization. In the case of doctoral students, this reigning in is by quite a lot, which highlights an important point: even if the total bias from TWFE is small, it is still important to use the most up to date identification techniques because there may be a sub group of the data for which the bias is large.

# 6 Conclusion

In this project, we replicate and extend Nagaraj and Reimers, 2021. Our extension serves mostly to confirm the validity of the original researchers' results, as the CS estimator does not flip the sign of any of the estimates produced by TWFE. However, the CS estimator does reign in the high magnitude estimate of the digitization ATT for doctoral student borrowers that TWFE produces, highlighting the point that even if the total bias from TWFE is small, it is still important to use the most up to date identification techniques because there may be a sub group of the data for which the bias is large. Like Nagaraj and Reimers, our work finds that the substitution effect dominates for library users at Harvard, likely due to the muting of the discovery effect in a setting where

search assistance is already easily available.

One thing our work highlights is the need to extend the recent work by Goodman-Bacon, Callaway and Santana, and Sun and Abraham to make clear how to estimate multiple ATTs with the same regression. To use the CS estimator, we were forced to filter our data by borrower type, instead of using dummy variables to shift our ATT between borrower groups as we do in TWFE. This extension to the recent work must also come with easy to implement support for researchers who use popular statistical software such as R and Stata

Being a platform, Google Books has a network effect, 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.

# References


Aguiar, L., & Waldfogel, J. (2018). As streaming reaches flood stage, does it stimulate or depress music sales? *International Journal of Industrial Organization*, *57*, 278–307.

Bounie, D., Bourreau, M., & Waelbroeck, P. (2006). Piracy and demands for films: Analysis of piracy behavior in french universities.

Callaway, B., & Sant'Anna, P. H. (2021). Difference-in-differences with multiple time periods. *Journal of Econometrics*, *225*(2), 200–230.

Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. *Journal of Econometrics*, *225*(2), 254–277.

Nagaraj, A., & Reimers, I. (2021). Digitization and the demand for physical works: Evidence from the google books project. *Available at SSRN 3339524*.

Rob, R., & Waldfogel, J. (2007). Piracy on the silver screen. *The Journal of Industrial Economics*, *55*(3), 379–395.

Smith, M. D., & Zentner, A. (2016). Internet effects on retail markets. In *Handbook on the economics of retailing and distribution.* Edward Elgar Publishing.

Sun, L., & Abraham, S. (2021). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics*, *225*(2), 175–199.