Online Appendix: Digitization and the Market for Physical Works: Evidence from the Google Books Project

Abhishek Nagaraj
University of California – Berkeley

Imke Reimers
Northeastern University

A Regression Discontinuity

The main analysis takes advantage of variation in the timing and status of digitization across all books in the Harvard Widener library system. An underlying assumption in these analyses is that books that are digitized are inherently similar to books that are not, or not yet, digitized. However, whether a book is digitized at all is a function of the book's copyright status. Throughout the time period of our study, all works that were originally published before 1923 are in the public domain and hence digitized, whereas works from 1923 and later were still protected by copyright and hence not digitized.

This discontinuity in copyright status is due to the most recent copyright extension in the United States. The 1998 Copyright Term Extension Act retroactively extended the copyright term for all protected works by twenty years, from 75 to 95 years for the works in our dataset. This extension provides a sharp, exogenous discontinuity in the digitization status for works originally published around 1923. Beyond their digitization status, however, nothing changed systematically for just one group of titles over the time period in our study. Thus, were it not for the digitization of the older works, one might expect analog demand for these titles to evolve similarly for works originally published on both sides of the 1923 cutoff.

We examine how demand changed from 2003/04 (before any works were digitized) to 2010/11 (after digitization was completed) in a regression discontinuity design that formalizes the patterns shown in Figure 3 in the main text. Formally, we utilize the jump in digitization around the original publication year 1923 to estimate regression equations of the form

$$Y_i = \alpha + \beta Digitized_i + k(year_i) + \epsilon_i, \tag{1}$$

where Y_i describes various measures of the change in book j's unit sales from 2003/04 to 2010/11, including the absolute unit changes, an asymptotic sine transformation of these changes, and an indicator that equals one if there is an increase in book i's demand.[1] Moreover, $Digitized_i$ is an indicator variable that is 1 if the book was digitized, which is a deterministic function of the book's original year of publication. We define $k(year_i)$ as a quadratic function of the book's publication year, centered around 1923. The bandwidth in each specification is its mean-squared-error optimal bandwidth.

Table E.8 shows the results from these specifications. All results support those in the main text: treatment through digitization leads to an increase in marketwide sales. The estimated percentage effects of around 28 percent are even larger than those from our main specifications. These results are robust to different bandwidths and functional forms of the publication year. Figure 3 Panel B further plots the annual coefficients of the regressions on the likelihood of increased demand (analog

[1] We use an asymptotic sine transformation instead of the more common log transformation because one would naturally expect many negative changes in demand, and dropping these may bias results (note that our dependent variable is the *change* in demand).

to column 3 in Table E.8). It illustrates that books originally published before 1923 and therefore digitized by Google Books are significantly more likely to experience an increase in sales.

B Amazon's Search Inside the Book Feature

Our main analysis has two major advantages: it relies on a clean natural experiment, and it allows us to quantify the effect of the full provision of digital texts on analog demand. Along with these advantages come some disadvantages. First, we only observe digitization of books that are in the public domain and therefore quite old. And second, the full-text digitization and access prevents the isolation of the discovery effect by muting the substitution effect. We address both concerns in a parallel analysis that focuses on Amazon's "Search Inside the Book" (SITB) program, which scans the entire text of books and makes them available for full-text search. However, the book itself is provided only in "snippet" view, so the SITB feature is unlikely to displace regular sales. In fact, authors and publishers can set the percentage of the book that customers can view between 10 and 80%, and the default is 20%.[2] If digitization-enabled search helps consumers discover new content, then books that adopt SITB are likely to have greater demand than books that do not adopt SITB.

Unlike Google Books, which was implemented en masse, the SITB feature is optional for publishers and authors to adopt.[3] An ideal experiment would randomly assign SITB status to a set of comparable books and examine the effect of this treatment on sales. Lacking such an experiment, we provide evidence from a cross-sectional fixed-effects specification for the hypothesis that SITB adoption raises sales. We start with a set of the top 1200 featured books in the "Business and Investing" and "Education and Reference" categories on Amazon. For these 2400 books, we obtain all books written by their authors. We thus build a sample of 1775 authors who have written a total of 11,166 books. Of these books, about 80 percent have the SITB feature enabled. While some of this variation is due to time (more recent books are more likely to have SITB enabled), there is considerable variation within years and authors.

Using this sample, we estimate the following specification: $Y_{ia} = \alpha + \beta SITB_{ia} + \gamma_a + \delta_i + \epsilon_{ia}$, where Y_{ia} indicates the total number of Amazon reviews or the natural log of the current bestseller rank for book i published by author a. $SITB_{ia}$ equals one if the book has SITB enabled, and zero otherwise. γ_a indicates author fixed effects and δ_i indicates publication year fixed effects. The key identification concern with this specification is that books with greater market potential are selectively SITB-enabled, while more marginal books are not. Controlling for author and year fixed effects helps considerably as we do not compare higher-selling authors vs. other authors, or publication years with greater demand vs. those with lower demand.

Table E.6 presents estimates from this analysis. The estimated demand effects are positive

[2] `https://www.amazon.com/gp/feature.html/ref=amb_link_6?ie=UTF8&docId=1001119751`

[3] This implies that the program also includes more recent books.

and significant. Compared to a base of 554 reviews, SITB-enabled books see about 214 more reviews (Col 2), an increase of about 38 percent. Similarly, SITB-enabled books have about a 73 percent lower (better) rank compared to books that are not part of the SITB program. Taken at face value, these results suggest that even for a sample of modern, in-copyright books, book digitization can significantly improve discovery of new content, thereby increasing demand. Note however that we cannot rule out the concern that authors selectively treat their better books with SITB. Nevertheless, if this is the case, a revealed-preferences argument would lead us to arrive at a similar conclusion: authors and publishers *want* their best books to be searchable.

C The Impact on Prices

The main text shows that digitization through Google Books leads to an increase in the number of editions, particularly from independent publishers. The digitization-induced increase in unit sales could be driven by a decrease in prices of the new editions. If such decreases in prices are large enough, then digitization could lead to decreases in revenues despite the increase in sales.

We obtain list prices from Bowker's Books-in-Print sample to examine the effect of digitization on the prices of new editions. We treat each newly published edition as an observation, and we estimate the edition's suggested retail price as a function of the title's digitization status at the time of the edition's publication. We further include indicators for the edition's year of publication and for each title. Formally, we estimate

$$Y_{jit} = \alpha + \beta PostScanned_{jit} + \gamma_i + \lambda_t + \epsilon_{jit}, \tag{2}$$

where Y_{jit} is the suggested retail price (or its log) of edition j of title i, which was published in year t. Moreover, $PostScanned_{jit}$ is an indicator that equals 1 if title i was digitized before the year t in which edition j was published, and γ_i and λ_t are title and publication year indicators, respectively. In additional checks, we add controls for the number of available editions for title i.

Table E.7 depicts the results from these regressions. None of the specifications show any evidence that Google's digitization program had an impact on prices of new editions, as all coefficients are small and statistically insignificant.

D Examining The First Stage: Digital Use of Google Books

The causal logic we outline in our study is that readers discover books on Google Books and a proportion of them end up buying a physical copy, thereby increasing sales. In order for this logic to hold, books digitized via Google Books must be read digitally and the online diffusion of content must be more pronounced for digitized books than non-digitized books. This section provides some suggestive evidence for this "first stage." Our goals are twofold: first, we want to gather data on

the overall use of Google Books for digitized titles and explore whether they receive a meaningful level of digital use; and second, we want to explore whether digital use is smaller for books that are not scanned at all, or are scanned but are not available in full-text format (i.e. in snippet or partial view).

Ideal data on this topic would come from Google's internal data on page-level traffic. Google Books is heavily used in that the overall website received almost 40 million visits in December 2021 according to SimilarWeb.[4] However, since traffic data is not available at the level of an individual Google Books URL, we turn to two proxies. First, we explore the citation behavior of Wikipedia articles to books (and to Google Books links) and second, we explore backlinks, i.e. HTTP links from third-party websites like blogs, newspapers etc. to Google Books pages on the wider web. Collectively, our data suggest that (a) books digitized by Google Books are used widely and (b) books available for full-text view are much more heavily used than those not available or available in partial view.

D.1 Wikipedia

To examine whether or not a book is cited on Wikipedia, we rely on data from Singh, West and Colavizza (2021), who extracted 29.3 million citations from 6.1 million English Wikipedia articles as of May 2020, and classified them as being books, journal articles, or Web content. We select all citations that have "books.google.com" in the URL or have an identifier that indicates it as a book (e.g. ISBN or OCLC). This process gives us a total of 2,587,149 citations.

D.1.1 Baseline Sample:

For the analysis presented in Figure 5 in the main text, we try to match the set of Wikipedia citations to the set of 88,006 books in our baseline sample. We do this by comparing the similarity ratio between the two title strings computed using the Levenshtein distance and choosing a threshold match ratio of 95%. We find that 4,295 books in our sample have at least one cite on Wikipedia when using this quite conservative threshold. That is, about 5% of books in our main sample are cited at least once.[5] Since citing a book on Wikipedia requires an editor to read the book and find it useful to cite, a 5% likelihood of citation offers a conservative estimate of the number of books that are being read through Google Books.

Figure 5 in the main text plots this likelihood of being cited by publication year, showing a sharp drop for books published after 1923. Since these books are also much less likely to be digitized, this pattern suggests that books in our sample that are not scanned are much less likely to be cited on Wikipedia. In fact, when using our data on scan status (rather than publication year) we find that scanned books are cited at the rate of 6.1%, while unscanned books are cited at the rate of 3.9%,

[4] `https://www.similarweb.com/website/books.google.com/`

[5] This number goes up to 10-12% when using less conservative thresholds, but we get a lot of false positive matches.

a difference of 2.2% percentage points. Note that our citations include both direct links to Google Books as well as purely ISBN references. Many post-1923 books have Google Books URLs with a limited view or no page at all. For these books, a Wikipedia editor could cite a link (if one exists) or cite the underlying ISBN of a book if there is no Google Books page. Even when considering these ISBN references, we find that pre-1923 books are much more likely to be cited when compared to post-1923 books. Not only do Wikipedians use Google Books, but their ability to reuse and cite it is hampered when books are not available in full scan view.

D.1.2 All Wikipedia Book Citations

We provide a similar analysis in In Figure E.5, except that we use the full sample of Wikipedia citations to Google Books links of books published between 1904 and 1943 (20 years before and after the 1923 cutoff), a set of 39,439 citations to 28,554 unique titles. For this sample, we explicitly compare titles by their access status as indicated by the Google Books API, and we plot the number of citations by year in Panel A. The results of this analysis provide an analog to our analysis using the baseline sample among a broader sample of Wikipedia citations. Conditional on a book being cited on Wikipedia, it is much more likely to come from pre-1923, publication years for which Google Books scans are available in full. Further, in Panel B of Figure E.5, we plot similar statistics but by the viewability status of the underlying work, which we obtain from the Google Books API. This panel shows that the number of Wikipedia citations is much higher for works with Full Access than for works with Partial or No Access. Under the assumption that the stock of books from each publication year (and for each access level) does not vary discontinuously around 1923, this is further evidence that Google Books access leads to more digital use.

Overall, our Wikipedia analysis strongly suggests that there is significant use for digitized Google Books within the Wikipedia editor community and access restrictions play a major role in limiting digital use. Citing a work is a high bar for digital use, and we expect that the number of works being read by this community is significantly higher. Therefore, the differences in digital use are likely to get even stronger if data on digital readership were available.

D.2 Semrush

As a second measure of digital use, we consider "backlinks" – links to a specific Google Books book URL from the wider internet. Such links could come not just from Wikipedia, but from the entire web including from blog posts, online forums, newspapers, etc. If a book is being read on Google Books, we expect a small proportion of readers to post the link online in order to share it with other readers (further stimulating digital readership).

To collect data on backlinks, we rely on `semrush.com` (Semrush, 2018), a web service that collects data on backlink activity for search engine optimization uses. Using the Google Books API, we identify the Google Books URLs for each of the books in our sample using fuzzy matching at

the title level. We are able to match 29,399 books to URLs through this process. For this sample, we obtain data on backlinks from Semrush. Similar to the Wikipedia analysis, we examine whether books published pre-1923 and those that are scanned are more likely to be backlinked than books published after 1923 and that are not scanned.

Figure E.6 plots the likelihood of a book having at least one backlink by publication year. As this figure shows, there is a sharp drop in this statistic after 1923, when Google Books are likely to be scanned only in partial view or snippet status. In fact, our data suggests that on average about 16.3% of URLs to titles published pre-1923 have at least one backlink, while this number drops to 8.8% post-1923. Scanned books are significantly more likely to be backlinked than books that are not scanned, and this difference is statistically significant.

Combined, both the Wikipedia and the backlinks analysis establish the existence of a first stage – that books digitized via Google Books are indeed read in digital channels, and that access restrictions lower the online diffusion of knowledge.

References

Semrush. 2018. "Semrush Data Extract." https://www.semrush.com/.

Singh, Harshdeep, Robert West, and Giovanni Colavizza. 2021. "Wikipedia citations: A comprehensive data set of citations with identifiers extracted from English Wikipedia." *Quantitative Science Studies*, 2(1): 1–19.

Sun, Liyang, and Sarah Abraham. 2021. "Estimating dynamic treatment effects in event studies with heterogeneous treatment effects." *Journal of Econometrics*, 225(2): 175–199.

E Appendix Tables and Figures

Table E.1: Varying the Constant Term in the Main Regressions

	Log-OLS (sales + constant)					
	(1) $+10^{0}$	(2) $+10^{-1}$	(3) $+10^{-2}$	(4) $+10^{-3}$	(5) $+10^{-4}$	(6) $+10^{-5}$
Post-scanned	0.0466	0.188	0.360	0.537	0.714	0.891
	(0.0130)	(0.0217)	(0.0319)	(0.0427)	(0.0536)	(0.0647)
Book FE	Yes	Yes	Yes	Yes	Yes	Yes
Year-location FE	Yes	Yes	Yes	Yes	Yes	Yes
N	82836	82836	82836	82836	82836	82836

Note: This table presents estimates from log-OLS models evaluating the overall impacts of book digitization on physical book sales. In each column, the dependent variable is $\ln(sales + \text{C})$, where C varies from 10^0 to 10^{-5}. Post-scanned equals one in all years after the book has been digitized. All columns include book and year-location fixed effects. Standard errors, clustered at the book level, are in parentheses.

Table E.2: Varying Effects by Scan Year

Panel A: Effects on Log-sales (OLS)

	Individual Scan Year				
	2005	2006	2007	2008	2009
Post-Scanned	0.0198	0.0751	0.0289	0.0531	0.171
	(0.0263)	(0.0243)	(0.0219)	(0.0226)	(0.0747)
Book FE	Yes	Yes	Yes	Yes	Yes
Year-Location FE	Yes	Yes	Yes	Yes	Yes
N	60480	60345	61353	60201	54261

Panel B: Effects on Any-sales (LPM)

	Individual Scan Year				
	2005	2006	2007	2008	2009
Post-Scanned	0.0675***	0.0875***	0.0804***	0.104***	0.154***
	(0.00907)	(0.00995)	(0.00891)	(0.0103)	(0.0372)
Book FE	Yes	Yes	Yes	Yes	Yes
Year-Location FE	Yes	Yes	Yes	Yes	Yes
N	60480	60345	61353	60201	54261

Note: This table shows the effect of digitization on sales across digitization cohorts. Panel A reports results from OLS estimations where the dependent variable is zero-inflated log-sales; Panel B reports estimates from linear probability models where the dependent variable is an indicator that is one if at least one copy of the book was sold. In both panels, the five columns report results from separate regressions for each digitization cohort. For example, in column (1) we keep all books scanned in 2005 as well as all unscanned books, and we estimate the effect of digitization in that year. We do the analogous exercise for subsequent scan years in columns (2) through (5). Post-scanned equals one in all years after the title has been digitized. All models include book and year-location fixed effects, and standard errors (in parentheses) are clustered at the book level.

Table E.3: Robustness Checks for Sales, by Popularity

	Endogeneity			Model		
	(1) Controls	(2) Pusey/2005	(3) Pusey/2005	(4) Asin(sales)	(5) Poisson	(6) Logit
Post-scanned:						
... × Presales=0	0.0458	0.0440	0.0457	0.0669	0.999	0.780
	(0.0127)	(0.0159)	(0.0158)	(0.0150)	(0.374)	(0.0895)
... × Presales>0	0.279	0.369	0.341	0.350	0.289	-3.634
	(0.101)	(0.139)	(0.135)	(0.115)	(0.110)	(0.369)
... × Presales>500	-0.222	-0.0570	-0.125	-0.148	0.293	9.060
	(0.101)	(0.126)	(0.127)	(0.106)	(0.156)	(427.8)
Book FE	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	No	No	No	No	Yes	Yes
Year-location FE	Yes	Yes	Yes	Yes	No	No
Addtl. Controls	Yes	No	Yes	No	No	No
N	82836	62199	62199	82836	26586	22671

Note: This table reports the robustness of the heterogeneity results. Columns (1) through (3) provide zero-inflated Log-OLS estimates that mirror columns (1) through (3) of Table 3. Column (1) adds controls for a book's Google Search volume as well as a dummy variable that equals one if the book has also been digitized on Project Gutenberg before year t. Column (2) drops all books digitized in 2005 or located in Pusey 3, and column (3) adds the demand controls from (1) to the sample from (2). Columns (4) through (6) vary the functional form, mirroring columns (7) through (9) in Table 3. Post-scanned equals one in all years after the book has been digitized. Presales=0 includes books with no sales in 2003 and 2004. Presales>0 describes books between 1 and 500 total sales in 2003 and 2004. And Presales>500 includes all books with more than 500 sales. All models include book and year fixed effects. Columns (1) through (4) additionally interact these year fixed effects with library-location fixed effects. Standard errors are in parentheses, clustered at the book level.

Table E.4: Sales Estimates with More Granular Popularity Groups

	Baseline		Robustness	
	(1) Log-Sales	(2) Log-Sales	(3) Public Domain	(4) Twins
Post-Scanned:				
... × Presales=0	0.0456***	0.0453***	0.0492***	0.0575***
	(0.0123)	(0.0128)	(0.0141)	(0.0165)
... × Presales>0	0.169	0.165	0.180	0.260
	(0.154)	(0.155)	(0.155)	(0.213)
... × Presales>40	0.481***	0.456***	0.460***	0.476***
	(0.132)	(0.133)	(0.133)	(0.182)
... × Presales>490	-0.0175	-0.0296	-0.0217	-0.127
	(0.239)	(0.237)	(0.238)	(0.313)
... × Presales>2020	-0.197	-0.213	-0.205	-0.0607
	(0.157)	(0.155)	(0.155)	(0.250)
... × Presales>7600	-0.0962	-0.131	-0.120	-0.0754
	(0.138)	(0.137)	(0.139)	(0.212)
Book FE	Yes	Yes	Yes	Yes
Year FE	Yes	No	No	No
Year-location FE	No	Yes	Yes	Yes
N	82836	82836	29394	36738

Note: This table provides robustness checks to our popularity cutoff choices. The table mirrors Panel A of Table 4. Presales=0 includes books with no sales in 2003 and 2004. Presales>0 describes books with 1 to 40 total sales in 2003 and 2004. Presales>40 includes all books with 41 to 490 sales in 2003 and 2004. And so on. That is, all popularity groups are mutually exclusive. All columns report results from zero-inflated Log-OLS regressions. Columns (1) and (2) report the baseline regressions. Columns (3) and (4) repeat the first two sample-related robustness checks from Table 3, using only public domain (digitized) books (3) and including only matched pairs (digitized and not) that are located exactly next to each other (4). Post-scanned equals one in all years after the book has been digitized. Column (1) includes book and year fixed effects. The other three columns use book and year-location fixed effects. Standard errors are in parentheses, clustered at the book level.

Table E.5: Robustness Checks for Loans Regressions

	Pre-05 Loans		Sales Sample		Model		
	(1) Log-Loans	(2) Any-Loans	(3) Log-Loans	(4) Any-Loans	(5) Asin(loans)	(6) Poisson	(7) Logit (0/1)
Post-Scanned	-0.0217	-0.123	-0.0251	-0.103	-0.0656	-0.484	-0.501
	(0.00201)	(0.00824)	(0.00202)	(0.00735)	(0.00196)	(0.0149)	(0.0122)
Book FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	No	No	No	No	No	Yes	Yes
Year-Location FE	Yes	Yes	Yes	Yes	Yes	No	No
N	287523	82836	287523	82836	792054	792054	791028

Note: This table examines the robustness of the loans regressions. Columns (1) and (2) restrict the sample to books that had at least one loan before the start of digitization (31,947 titles). Columns (3) and (4) estimate the effect on loans using only those books for which we also have sales data (9,204 titles). Columns (5-7) provide estimates using alternate functional form assumptions. Log-Loans describes OLS regressions where the dependent variable is the zero-inflated log of loans. Any-Loans describes linear probability models where the dependent variable equals 1 if at least one copy of the book is sold in year t. Asin(Loans) represents the inverse hyperbolic sine transformation of the Loans variable. Post-Scanned equals one in all years after a book has been digitized. Book and year-location fixed effects are included in all models. Standard errors are in parentheses, clustered at the book level.

Table E.6: Muting the Substitution Mechanism: Amazon Search Inside the Book

	Public Domain Sample		Modern Sample	
	(1) Reviews	(2) Reviews	(3) Reviews	(4) Reviews
SITB Enabled	269.1	214.4	-0.781	-0.739
	(57.10)	(65.60)	(0.0415)	(0.0467)
Author FE	Yes	Yes	Yes	Yes
Book FE	—	—	—	—
Year FE	No	Yes	No	Yes
N	11127	10572	10826	10344

Note: This table examines the effect of Amazon's "Search Inside the Book" (SITB) scheme, which scans the entire contents of a book and permits readers to search the full text of the book, but does not allow them to read the entire text. We collected a sample of 11,166 in-copyright books by 1775 authors. We examine the effect of SITB in a cross-sectional specification regressing the cumulative number of reviews (Cols 1-2) and ln(Rank) (Cols 3-4) as of Dec 2020 on SITB status, with author fixed effects and/or release year fixed effects (Cols 2,4).

Table E.7: Impact of Digitization on Prices of New Editions

	OLS		Log OLS	
	(1) Price	(2) Price	(3) Ln(Price)	(4) Ln(Price)
Post-Scanned	0.181	0.268	-0.00901	-0.00681
	(0.821)	(0.826)	(0.00923)	(0.00929)
Editions		0.0226***		0.000572***
		(0.00478)		(0.0000862)
Book FE	Yes	Yes	Yes	Yes
Year-location FE	Yes	Yes	Yes	Yes
N	275395	275395	275270	275270

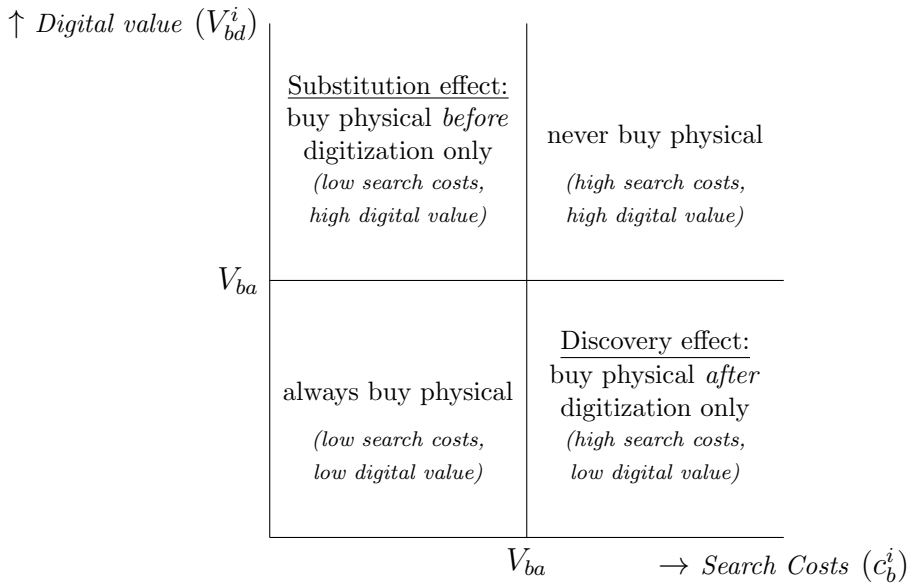
Note: This table presents estimates from OLS and log-OLS models evaluating the impact of book digitization on list prices of newly published editions. The estimation is on the edition level, including all 275,395 editions published between 2003 and 2011 that we matched between the titles in Harvard's library system and the Bowker Books-in-Print directory. Post-Scanned equals one in all years after a book has been digitized. Book and year-location fixed effects are included in all models. Standard errors are in parentheses, clustered at the title level.

Table E.8: Regression Discontinuity Estimates

	Sales		
	(1) sales	(2) asinh(sales)	(3) increase
Digitized	577.4	0.277	0.159
	(450.8)	(0.126)	(0.0229)
N	8016	8016	8016

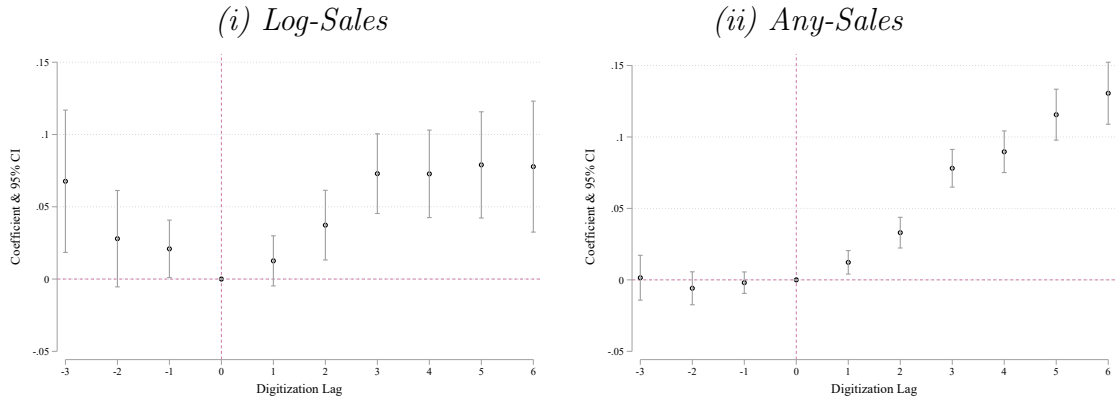
Note: This table presents results from regression discontinuity estimations on the title level. The dependent variables are functions of the changes in analog sales between 2003/04 (before digitization) and 2010/11 (after digitization). In column (1), it is the absolute change in analog sales; column (2) uses the asymptotic sine of that change; and column (3) uses an indicator that is 1 if analog demand has increased. The independent variable of interest, *Digitized*, is an indicator that is 1 if the book was digitized (i.e. originally published before 1923). A quadratic function of the publication year is included. The bandwidth in each specification is the MSE-optimal bandwidth. Robust standard errors in parentheses.

Figure E.1: Theoretical Framework: Decision to Consume Physical vs. Digital



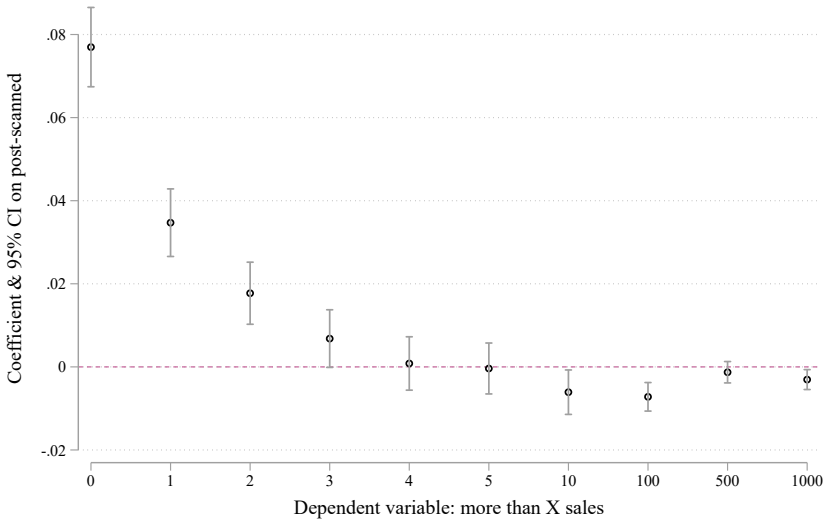
Note: This figure provides an illustration of predictions from the theoretical framework. The framework models an individual customer i's decision to purchase an analog version of the book (a physical copy) as a function of his or her search costs c_b^i (x-axis) and their valuation of the digital copy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

Figure E.2: Event Study Plots Adjusted for Heterogeneous Treatment Effects



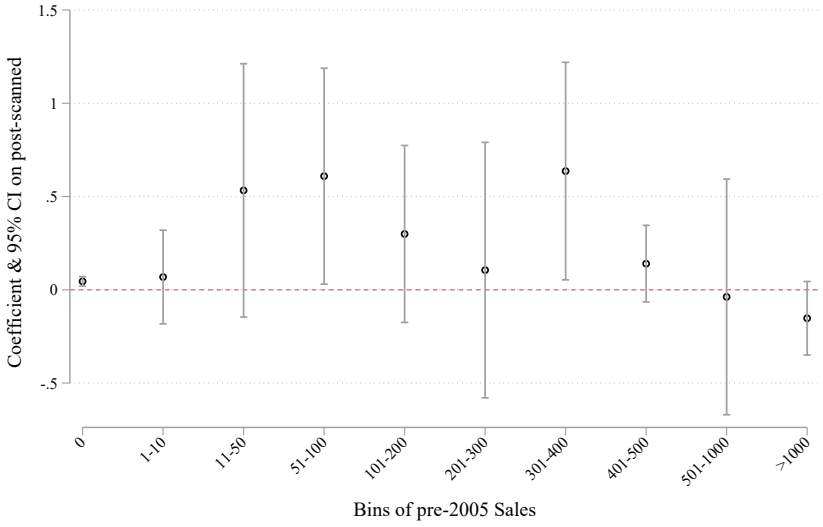
Note: This figure presents event study plots similar to Figure 4 in the main text, adjusting for the possibility of heterogeneous treatment effects in a setting where books are treated in different cohorts at different points in time. The event study estimates are adjusted using the Sun and Abraham (2021) estimator.

Figure E.3: Estimated Probability of Surpassing Varying Sales Thresholds



Note: This figure shows estimated coefficients and their 95% confidence intervals of the "post-scanned" variable in separate regressions. The dependent variable in each regression is an indicator that equals one if the sales threshold indicated on the x-axis is surpassed for a book in a year. The "post-scanned" variable equals one in all years after the book has been digitized. The regression includes book and year-location fixed effects, and standard errors are clustered at the book level.

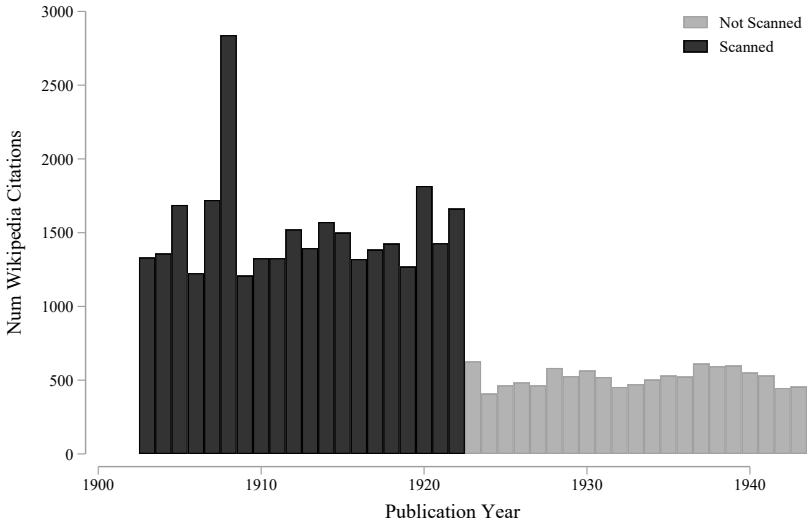
Figure E.4: Sales Estimates with Granular Popularity Groups



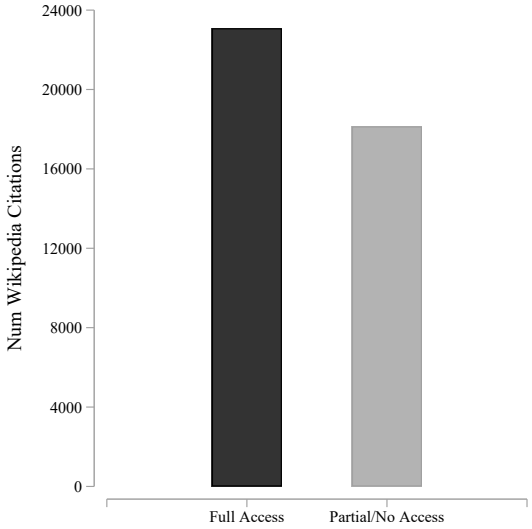
Note: This figure reports coefficients and 95% confidence intervals of the interactions of the "post-scanned" variable with mutually exclusive groups of pre-2005 sales, from a zero-inflated Log-OLS regression. The "post-scanned" variable equals one in all years after the book has been digitized. The regression includes book and year-location fixed effects, and standard errors are clustered at the book level.

Figure E.5: Citations to Scanned and Unscanned Works on Wikipedia (1904-1943)

A. By Year

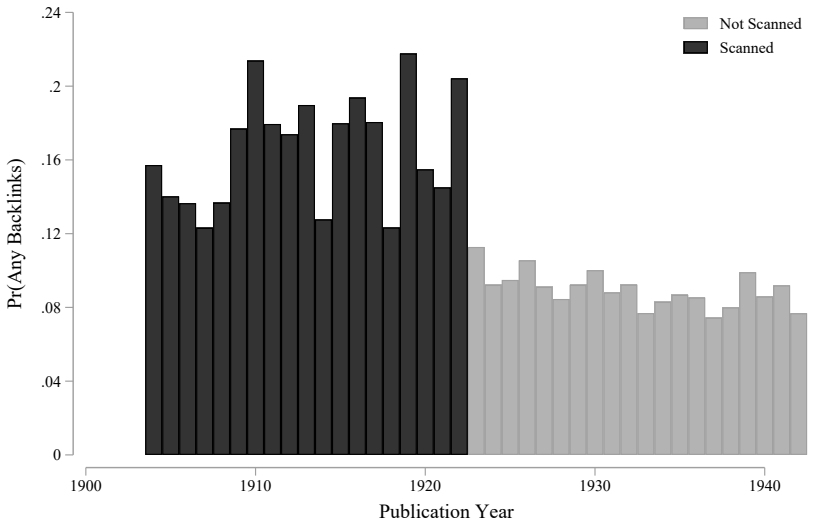


B. By View Status



Note: This figure presents data from the full sample of Wikipedia citations to books published between 1904 and 1942. This sample consists of 39,439 citations to 28,554 unique titles. For each title, we determine its viewability status (full access, partial access or no access) by looking for the access characteristics of the same title via the Google Books API. We present counts of the number of citations by publication year (Panel A) and by scan status (Panel B).

Figure E.6: Backlinks to Scanned and Unscanned Works on Google Books (1904-1943)



Note: This figure presents data collected from semrush.com, which provides information on links from the web to specific books pages on Google Books for the 29,399 books in our sample published between 1904 and 1942 that we can match with high confidence. We present the average likelihood that a title has at least one backlink by publication year.