Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Does Open Access Matter for Citations? A Topic-Stratified Analysis of Library and Information Science Discipline

Arindom Kalita
Mizoram University
Kumar Sumit
aryan.sumit121@gmail.com
Mizoram University
Manendra Singh
Mizoram University
Divyanshu Gupta
Mizoram University

Research Article

Keywords: Open access, Citation advantage, Topic modeling, Citation analysis, Bibliometrics, Regression analysis

Posted Date: August 31st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10855106/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Does Open Access Matter for Citations? A Topic-Stratified Analysis of Library and Information Science Discipline

Arindom Kalita[1†], Kumar Sumit[1*†], Manendra Singh[1†], Divyanshu Gupta[1†]

[1]Department of Library and Information Science , Mizoram University, Tanhrill, Aizawl, 796004, Mizoram, India.

*Corresponding author(s). E-mail(s): aryan.sumit121@gmail.com;
Contributing authors: kalitaari8@gmail.com; manebhu007@gmail.com; divyanshu6g@gmail.com;
†These authors contributed equally to this work.

Abstract

Open access publishing aims to make research accessible to a wide range of readers, with broader accessibility offering a potential pathway to increased citation exposure. Citation impact therefore remains an important justification for OA policies and motivates the question examined in this study. Using BERTopic, 52 research topics were derived from 14,720 Scopus-indexed Library and Information Science (LIS) publications published between 2010 and 2024. This study examined the OA–citation association overall, within individual research topics, and across progressively controlled regression models. OA publications showed a modest overall citation association ($\boldsymbol{r = .141}$, $\boldsymbol{p < .001}$). This association was not uniform across topics: among the 50 topics included in the topic-level analysis, 40 (80%) showed no detectable OA association, while 10 topics, concentrated in health-sciences-adjacent LIS subfields, accounted for most of the aggregate association. In the fully specified regression model, the journal-level citation baseline was the strongest predictor of citation rate ($\boldsymbol{\beta = .235}$, $\boldsymbol{p < .001}$), exceeding the coefficients for international collaboration, document type, and author count. Its inclusion reduced the OA coefficient by more than half, from $\boldsymbol{\beta = .114}$ to $\boldsymbol{\beta = .054}$ (both $\boldsymbol{p < .001}$). There was no statistically detectable temporal change in the OA–citation association from 2010 to 2022. Findings indicate that the citation association of OA is topic-dependent and conditional on the journal-level citation environment. This study uses an observational design and therefore estimates statistical associations rather than causal effects.

Keywords: Open access; Citation advantage; Topic modeling; Citation analysis; Bibliometrics; Regression analysis

Introduction

Open access (OA) publishing has moved, over the past decade, from a voluntary alternative into an increasingly codified policy requirement across major research funders, from Plan S's zero embargo principle to the EU's Horizon Europe mandate, the 2022 US OSTP memorandum, and comparable mandates in the UK, Canada, Japan, and South Africa. The longitudinal expansion of open-access publishing over this period has also been documented in bibliometric analyses of the OA landscape [1]. Among the arguments most frequently invoked in support of these mandates is the claim that open access increases a paper's citation impact, the open access citation advantage, or OACA. This claim has a substantial empirical history but also a persistent methodological weak point that motivates the present study. To establish the empirical and methodological context for the present study, a focused literature search was conducted to identify relevant studies of the open-access citation association and its methodological determinants. The identification, screening, and thematic categorisation of the included studies are shown in Fig. 1.

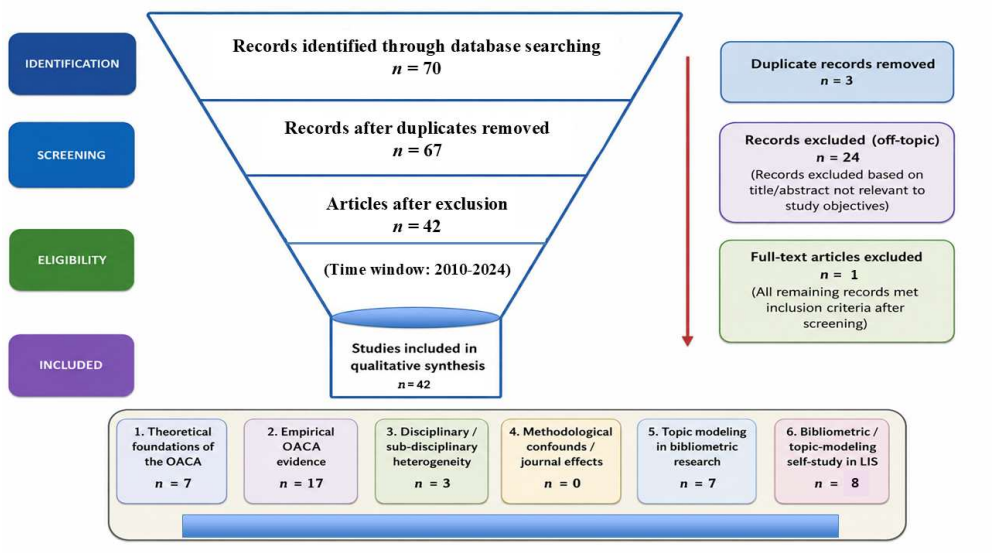


Fig. 1 Literature identification, screening, and thematic categorisation of studies included in the review.

The earliest tests of the OACA, conducted on arXiv preprints in physics and mathematics, could not agree even with one another on the underlying mechanism. Moed [2] decomposed the citation advantage of condensed matter physics preprints into an early view effect and a quality bias effect, concluding that early view accounted for most of the advantage in that field. Davis and Fromerth [3], examining mathematics, found a 35% citation advantage for arXiv deposited articles but little support for a universal open access explanation and no empirical support for early view, attributing the advantage instead to a quality differential. Gargouri et al. [4] compared self selected against mandated deposition, reasoning that it removes an author's ability to selectively archive only stronger work, and found the advantage persisted under both conditions, interpreting this as evidence for a genuine accessibility effect, though persistence under mandate and causal accessibility remain conceptually distinct claims. Craig et al. [5] offered an early critical review and noted that studies of this period frequently implied causality without adequate consideration of confounding factors, a critique that has aged well. A systematic review roughly fifteen years later [6] examined over 100 OACA studies and found a majority reported a positive association, but identified persistent inconsistency in citation window definition, OA type classification, and, most consequentially here, confound handling. A substantial share of OACA studies compare OA and non-OA citation counts without controlling for the journal in which a paper appears. Journal level factors such as indexing reach, readership size, and editorial selectivity being well established predictors of citation performance in their own right. Other large-scale analyses have likewise examined OA in relation to citation patterns and the characteristics of the citations received, reinforcing the need to distinguish citation volume from broader dimensions of research impact [7].

Complementing these citation side accounts, Miguel et al. [8] reframed the question as one of visibility rather than citation counts alone, mapping the proliferation of open access journals across roughly 17,000 Scopus indexed titles and arguing that OA's primary effect is on a work's discoverability, of which citations are only one downstream indicator. Warlick and Vaughan [9] approached the same phenomenon from the supply side, interviewing biomedical faculty about their motivations for choosing open access venues and finding that perceived visibility and broader readership, rather than citation maximizing strategy, were the dominant stated motivations. Holmberg et al. [10] extended the visibility framework to altmetric attention, finding a comparable OA linked advantage that varied by discipline and platform, an early signal of the disciplinary heterogeneity this study investigates at finer resolution.

That heterogeneity has since been documented more directly. The OACA is generally reported as stronger in STEM fields than in the social sciences and humanities [11], and recent discipline specific studies extend this variation to individual subfields. Clayson et al. [12] found a robust advantage in human electrophysiology; Clark et al. [13] noted the benefit of paying for open access varied considerably across biology subfields; Fleming and Cook [14] and Friederichs [15] each found evidence of an advantage in special education and medical education respectively and Basson et al. [16] showed the advantage's existence and size depended on both the subject area examined and the specific citation metric used to define it. Not every discipline-specific test

finds a positive result: Lin [17], examining 24,176 articles across energy and fuels journals, found that the OA mechanism influenced publication lag but did not produce a detectable citation advantage, underscoring that the heterogeneity documented here includes null findings, not only variation in magnitude. Two studies bear directly on the present setting. Khan et al. [18] compared OA and non-OA journals within library and information science and found OA journals achieved greater citation impact, and Lee and Shim [19], using 109,759 LIS articles from 2001 to 2024, examined OACA variation by time, journal quartile, and OA type within the discipline. Both confirm an aggregate OACA in LIS but neither decomposes the discipline into its constituent research topics or incorporates an explicit journal level control.

A further cluster of work has isolated single, narrower confounds without addressing venue level effects comprehensively. Dorta-González and Dorta-González [20] found that OA articles in hybrid journals received disproportionately higher citations specifically when the underlying research was also funded, direct evidence of a funding confound. Momeni et al. [21], studying over 500,000 Springer Nature articles, found OA publishing itself was systematically associated with author country income level. Borrego [22] found a relationship between high impact metrics and higher, faster rising article processing charges. Maddi and Sapinho [23] described a Sci-Hub paradox in which widespread shadow library access may itself have eroded any measurable OACA by making non-OA articles nearly as accessible in practice as OA ones, a possibility Perry and Pereira [24] tested directly by comparing citation advantages before and after Sci-Hub's 2021 cessation of updates. A related strand distinguishes open access from other, frequently co-occurring open-science practices: Colavizza et al. [25], analyzing roughly 122,000 PLOS publications, found that preprint deposition and open data sharing were each independently associated with citation advantages of around 20% and 4%, respectively, while sharing code alone showed no detectable effect, and Pan et al. [26], using a multi-layered causal inference design on psychology articles, found open data sharing was associated with a 20% citation increase that was especially pronounced for non-open-access, unfunded, and internationally collaborative work. These findings suggest that at least part of what is attributed to open access status in less carefully specified studies may be confounded with correlated open-science practices rather than access status itself. A dedicated search for studies employing journal fixed effects, multilevel citation models, or matched designs to isolate OA status from venue level effects returned no directly relevant results in the literature reviewed for this study, consistent with a pattern in which individual confounds are studied in isolation but a comprehensive, journal level solution has not converged.

Topic modeling offers a natural tool for testing whether this kind of heterogeneity exists within a single discipline, since it identifies a corpus's thematic structure directly from publication text rather than from researcher imposed subject categories. The approach is well established for descriptive mapping of a field's themes, including within LIS itself: Wijewickrema [27], Siddique and colleagues [28, 29], and Mokhtari et al. [30] have each mapped LIS's bibliometric landscape, while related LIS bibliometric studies have examined themes such as the mapping of artificial intelligence research in academic libraries [31]. Gao et al. [32] and Wang et al. [33] have applied dynamic and BERTopic based topic models respectively to bibliometric corpora, the latter using

the identical framework adopted here. Applications of topic modeling in other research domains, likewise demonstrate its use for identifying research themes and their evolution, including marketing [34], advanced manufacturing [35], hospitality and tourism [36], personalized language learning [37], artificial intelligence in education [38], smart learning [39], and the educational research on the metaverse [40]. Velez-Estevez et al. [41] additionally found that internationally co-authored LIS papers receive more citations than domestic ones, a structural finding this study's own analysis independently corroborates. Despite this, to our knowledge, existing applications of topic modeling in this literature have primarily focused on descriptive mapping rather than using topic structure as a stratifying variable in subsequent citation-impact analysis. The present study extends this literature by combining topic-level stratification with a journal-controlled OACA design.

To our knowledge, existing studies have not combined a data-driven topic modeling decomposition of a single discipline with a citation-impact regression design to examine internal heterogeneity systematically. The literature also provides limited evidence on the role of journal-level citation environment within such a design, despite repeated calls for more consistent treatment of potentially influential factors in OACA research. Library and Information Science, as a field closely concerned with scholarly communication and bibliometrics, provides a well-suited setting for examining both dimensions simultaneously.

This study pursues three objectives:

- **RO1:** To estimate the association between open access status and citation rate in LIS and assess how it changes after adjustment for publication year, author count, international collaboration, document type, and journal-level citation environment.
- **RO2:** To determine whether the association between open access status and citation rate varies across research topics within LIS and identify topics in which the association is statistically detectable.
- **RO3:** To assess whether the association between open access status and citation rate varies over time between 2010 and 2022.

These objectives correspond to three research questions.

- **RQ1:** Does open access status show a statistical association with citation rate in LIS literature, and does this association persist after accounting for author team size, international collaboration, document type, and journal level citation norms?
- **RQ2:** Is this association, if present, uniform across the discipline's constituent research topics, or is it concentrated in identifiable subfields?
- **RQ3:** Has the magnitude of this association changed over the period 2010 to 2022, characterized by substantial growth in OA publishing and policy adoption?

Materials and methods

The overall analytical workflow is summarized in Figure 2. The study proceeds from corpus construction and data cleaning through topic modeling and robustness validation, followed by derivation of bibliometric and citation-based variables and inferential statistical analysis.

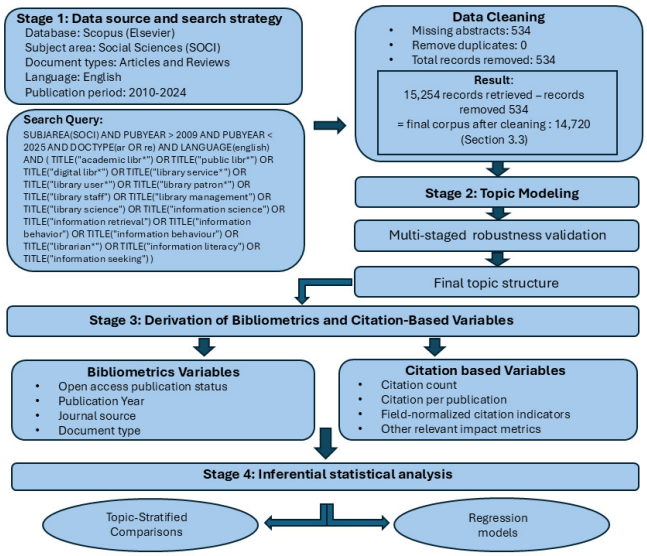


Fig. 2 Overview of the study workflow, from Scopus data retrieval and cleaning through topic modeling, variable derivation, and inferential statistical analysis.

Research design

This study combines topic modeling with inferential statistical analysis to examine the relationship between open access status and citation performance in Library and Information Science literature, and to test whether this relationship varies across the discipline's constituent research topics. The analysis proceeds in four stages: corpus construction and preprocessing, topic modeling with multi-stage robustness validation, derivation of citation based variables, and statistical analysis, including topic stratified comparisons and progressively controlled regression models.

Data source and search strategy

Bibliographic records were retrieved from Scopus database, chosen for its structured metadata, citation count reporting, and open access indexing. It has also been used as a bibliometric source for examining publication practices and journal-level patterns across countries and fields [42]. The search was restricted to the Social Sciences subject area, articles and reviews only, English language, and publication years 2010 to 2024, with keyword anchoring restricted to LIS specific terminology after an initial broader query returned substantial contamination from unrelated fields using bibliometric methods (see Figure 1). The final query returned 15,254 records and source title inspection confirmed concentration in recognized LIS venues.

Data cleaning

A total of 534 records were identified and excluded from the corpus due to missing or placeholder abstract text. Bracketed non-English original language titles were stripped

using regular expression matching, and title and abstract fields were concatenated to form the document unit for topic modeling. After these exclusions, the final analytic corpus comprised 14,720 records.

Topic modeling

Topic modeling was performed using BERTopic [43], combining transformer based document embeddings, dimensionality reduction, density based clustering, and class based TF-IDF topic representation. The primary embedding model was all-MiniLM-L6-v2. Dimensionality reduction used UMAP with 15 neighbors, 5 components, and cosine distance; clustering used HDBSCAN with excess of mass cluster selection; topic representations used a class based TF-IDF vectorizer with English stopword removal and unigram to bigram tokenization.

The HDBSCAN minimum cluster size parameter was selected through a sweep across candidate values from 15 to 100, jointly optimizing topic coherence and outlier proportion. A value of 40 produced the best balance of the two, coherence statistically indistinguishable from the sweep maximum at substantially fewer unclassified documents. Unclassified documents, 29.3 percent of the corpus, were reassigned to their nearest topic using BERTopic's outlier reduction strategy. The final model comprised 53 topics. Unclassified documents were excluded, yielding 52 substantive topics with a final coherence score of 0.631. Topics were labeled manually by the authors based on the top keywords and most representative documents per topic.

Three robustness checks were conducted. A higher capacity embedding model was re-run, all-mpnet-base-v2, yielded 79 topics with a mean best match Jaccard similarity of 0.714 against the primary model. A third, citation graph trained model, SPECTER2, produced markedly coarser topic granularity, 5 to 9 topics, and was excluded from further comparison. Re-fitting the primary model on five bootstrap resamples of 80 percent of the corpus yielded stable topic counts, 47 to 55 against 53 in the full corpus, and moderate keyword level stability, mean Jaccard 0.439, with instability concentrated in smaller topics and near perfect stability among the largest ones. The topic structure was also checked for plausibility. Its largest topics match well-established LIS subfields, each with its own journals, conference tracks, and professional associations, rather than looking like random artifacts of this particular corpus or embedding model.

Variable operationalization

Open access status was derived from the Scopus Open Access field and dichotomized, coded 1 for any indexed OA type and 0 otherwise. Citation rate was calculated as

$$\textit{Citation rate} = \frac{\textit{Citation count}}{\textit{Citation age}} \tag{1}$$

where citation age is the number of full calendar years between a paper's publication year and 2026, the year in which citation counts were retrieved from Scopus and fixed for this analysis. This reference point is distinct from 2024, the final publication year included in the corpus, since papers published as late as 2024 continued to accumulate citations up to the point of data retrieval. No partial year adjustment within

a calendar year was applied. The resulting rate was log transformed for regression analysis to address its right skew, using

$$\textit{Log citation rate} = \ln(1 + \textit{Citation rate}) \tag{2}$$

to accommodate papers with zero citations. Equation 1 defines the raw citation rate used throughout this study, and Equation 2 defines its log-transformed counterpart used specifically in the regression models reported in the Results. A citation analysis subsample restricted to 2010 to 2022 publications, 12,565 records, excluded 2023 and 2024 publications to allow sufficient citation accumulation time.

Author count, international collaboration, coded as a binary indicator of author affiliations spanning more than one country, and document type, article versus review, were derived from the corresponding Scopus metadata fields.A journal-level citation baseline was constructed for each paper i as the mean citation rate of all other papers in the same source title within the citation analysis subsample, excluding paper i itself:

$$\textit{Journal baseline}_i = \frac{1}{N_j - 1} \sum_{k \neq i} \textit{Citation rate}_k \tag{3}$$

where j denotes the source title containing paper i, and N_j is the number of papers from that source title in the citation analysis subsample. This leave one out construction avoids the circularity of relating a paper's citation rate to a journal average that includes that same paper. The baseline was restricted to source titles contributing at least 10 papers to the citation analysis subsample, since leave one out means computed from very few papers are unstable; this excluded 1,471 records from journals below that threshold, yielding a final analytic sample of 11,094 for models incorporating this covariate. The baseline is a within corpus behavioral measure rather than an independently validated prestige indicator such as CiteScore or SJR.

Statistical analysis

The overall association between OA status and citation rate was assessed using the Mann-Whitney U test with rank biserial effect size, given the pronounced right skew of citation count data. This comparison was repeated independently within each topic, restricted to topics in which both the OA and non-OA subgroups contained at least 5 papers. Two of the 52 topics fell below this threshold and were excluded, leaving 50 topics for stratified testing. Benjamini-Hochberg false discovery rate correction [44] was applied across these 50 tests, supplemented by a two way analysis of variance testing the OA by topic interaction on log transformed citation rates.

A sequence of ordinary least squares regression models was estimated with heteroscedasticity-consistent standard errors. Covariates were introduced progressively: publication year, then author count and international collaboration, then document type and the journal-level citation baseline, arriving at a fully specified model. A supplementary model added an OA by year interaction term to test temporal stability. An additional model restricted to the ten topics with the highest cumulative citation volume incorporated a full OA by topic interaction term within a single specification. Two further checks, addressing different design choices, assessed robustness:

a negative binomial regression of raw citation counts with log citation age as an exposure offset, and a re-run of the topic stratified analysis on the pre-reassignment cluster structure, excluding rather than reassigning HDBSCAN's unclassified documents. As a further robustness check on the journal-level citation baseline specifically, the fully specified model was re-estimated twice more: once replacing the continuous baseline with a full set of journal fixed effects, and once as a multilevel model with random intercepts for journal. These specifications test whether the continuous, leave-one-out baseline used as the primary covariate produces a materially different OA coefficient than more conservative approaches to journal-level clustering.

This is an observational bibliometric study based on publication level records. Open access status was not randomly assigned, and no instrumental variable, regression discontinuity, or other quasi-experimental strategy was used to isolate a causal effect. Reported associations and regression coefficients, referred to as OA coefficients throughout, should be interpreted as conditional statistical associations net of the measured covariates rather than causal estimates.

Analyses were conducted in Python using pandas, scipy.stats, statsmodels, scikit-learn, bertopic, sentence-transformers, umap-learn, hdbscan, and gensim. Figures were produced using matplotlib and seaborn.

Results

Corpus Characteristics and Topic Structure

The final analytic corpus comprised 14,720 LIS publications indexed in Scopus between 2010 and 2024. Publication volume increased over the study period, accompanied by a substantial increase in the proportion of publications carrying an indexed open-access designation, consistent with the broader growth in open access prevalence documented in large-scale, multi-disciplinary analyses [45]. Figure 3 presents these two temporal patterns together and establishes the publication and OA context of the corpus used for the subsequent analyses.

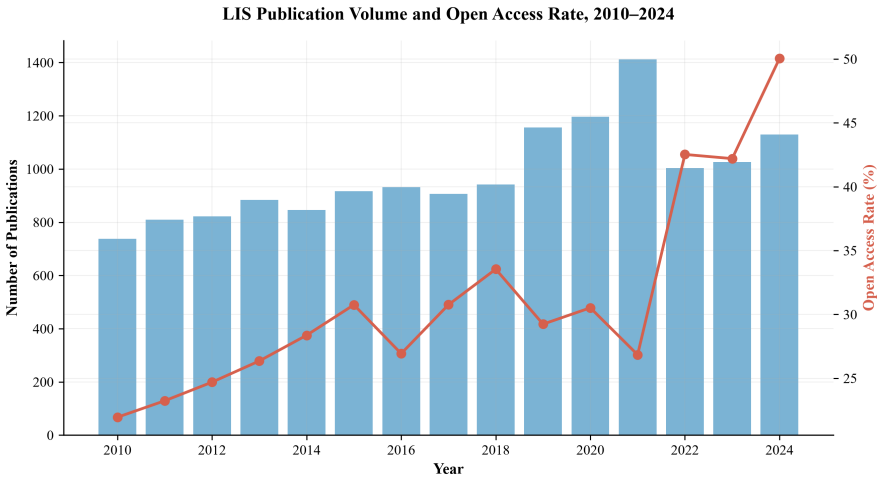


Fig. 3 LIS publication volume and open-access rate, 2010–2024.

The BERTopic analysis identified 52 substantive research topics from the final corpus. The topics covered a broad range of LIS research, including information literacy, public and academic libraries, information-seeking behaviour, bibliometrics and scholarly communication, information retrieval, digital libraries, health information, and library management. To provide a substantive view of the resulting topic structure, Table 1 presents the fifteen largest topics together with their document counts, shares of the corpus, and representative BERTopic terms.

Table 1 Fifteen Largest Research Topics and Representative BERTopic Terms

Rank	Research topic	n	%	Representative terms
1	Information Literacy Instruction	2,344	15.92	information literacy; literacy; information; students; IL; learning; instruction; skills; teaching; school
2	Public Libraries – Community & Value	977	6.64	public; public libraries; public library; libraries; library; social; community; services; value; study
3	LIS Education & Practice (Developing Regions)	855	5.81	library; study; Nigeria; libraries; ICT; university; librarians; resources; services; LIS
4	Information Seeking Behavior (General)	810	5.50	seeking; information; information seeking; behaviour; behavior; information behaviour; study; research; seeking behaviour; information behavior
5	Health Information Seeking Behavior	791	5.37	health; seeking; information seeking; health information; information; cancer; risk; COVID; COVID-19
6	Academic Librarians & Faculty Status	749	5.09	librarians; academic; research; faculty; librarian; academic librarians; professional; library; librarianship; job
7	LIS Bibliometrics & Scholarly Communication	736	5.00	LIS; science; journals; information science; research; library information; articles; journal; citation; authors
8	Library Collections & E-books	482	3.27	books; collection; book; academic; libraries; academic libraries; collections; library; print; electronic
9	Information Retrieval Systems	415	2.82	retrieval; information retrieval; IR; query; search; based; systems; information; documents; semantic
10	Digital Libraries & Preservation	392	2.66	digital; digital library; digital libraries; library; libraries; preservation; collections; archives; metadata; heritage
11	Librarian Job Satisfaction & Motivation	336	2.28	job; librarians; job satisfaction; knowledge; satisfaction; performance; work; study; management; stress
12	Academic Library Strategic Planning	330	2.24	academic; academic libraries; libraries; change; strategic; library; planning; academic library; management; future
13	Embedded/Distance Learning Librarianship	326	2.21	embedded; distance; online; learning; embedded librarianship; librarian; librarians; students; librarianship; embedded librarian
14	Health Sciences & Clinical Librarianship	307	2.09	health; medical; health sciences; sciences; clinical; librarians; hospital; librarian; care; sciences librarians
15	COVID-19 & Library Services	296	2.01	COVID; COVID-19; pandemic; library; libraries; services; academic libraries; librarians; information; health

The distribution of topic sizes was highly uneven. Information Literacy Instruction was the largest topic, comprising 2,344 publications (15.92%), followed by Public Libraries – Community & Value (977; 6.64%) and LIS Education & Practice (Developing Regions) (855; 5.81%). The fifteenth-largest topic, COVID-19 & Library Services, comprised 296 publications (2.01%). Figure 4 visualizes the relative size of these fifteen largest topics. The uneven distribution is relevant to the subsequent topic-stratified analysis because the empirical comparisons were conducted across research topics that differed substantially in their corpus representation.

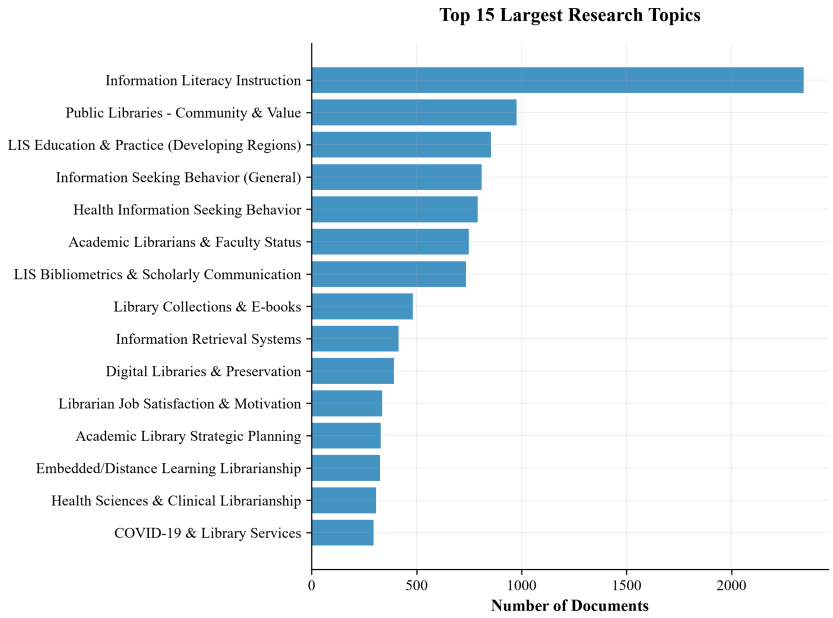


Fig. 4 Fifteen largest research topics by number of documents.

Overall Open-Access Citation Association

Having established the structure of the corpus and its topic distribution, the study next examined whether open-access publications differed from non-open-access publications in citation rate at the aggregate level. Open-access publications had a higher mean citation rate than non-open-access publications (2.09 versus 1.53 citations per year), with corresponding medians of 1.00 and 0.71. The Mann–Whitney U test indicated a statistically significant difference between the two distributions ($U = 18,606,116.5$, $p < .001$), with a rank-biserial effect size of $r = .141$. Figure 5 displays the distributions underlying this comparison.

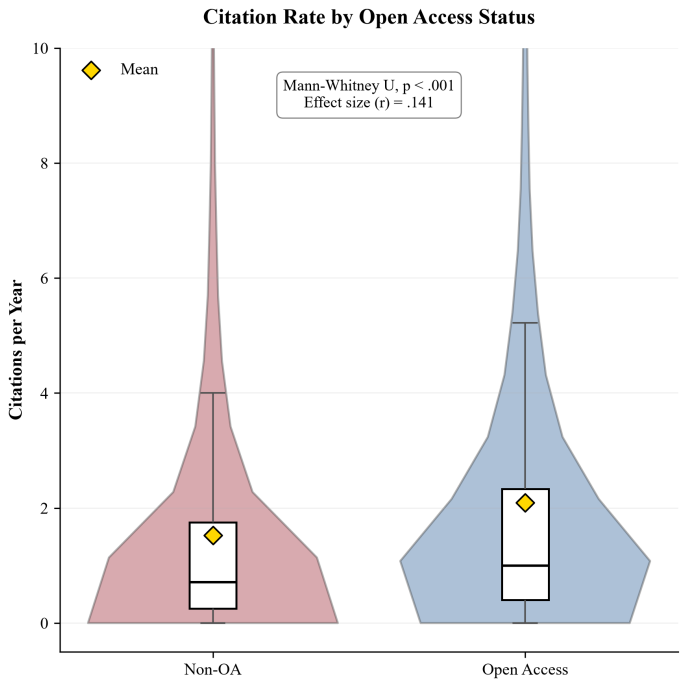


Fig. 5 Citation rate by open-access status.

Both distributions were strongly right-skewed, as expected for citation data, but the OA distribution was shifted upward relative to the non-OA distribution. The higher median as well as the higher mean indicates that the aggregate association was not attributable solely to a small number of highly cited OA publications. This claim was tested directly: open access publications were significantly overrepresented among the most highly cited outliers in the citation analysis subsample, with 46.8 percent of the top 1 percent most-cited papers being open access compared with 29.2 percent open access overall ($\chi^2 = 18.4$, $p < .001$). Despite this overrepresentation, the overall association held under progressively more stringent treatment of these outliers (Table 2).

Table 2 Sensitivity of the Overall OA-Citation Association to Highly Cited Outliers

Specification	**Effect size (r)**	p	n
Full sample	0.141	$< .001$	12,565
Excluding top 1% most-cited	0.136	$< .001$	12,439
Excluding top 5% most-cited	0.134	$< .001$	11,935
Winsorized at 99th percentile	0.141	$< .001$	12,565

At the topic level, excluding the five most highly cited papers within Health Sciences & Clinical Librarianship, the topic carrying the strongest association and among the smallest significant subgroups, left its effect size nearly unchanged ($r = .467$ versus $r = .475$ in the full topic). These checks indicate that open access papers do tend to include a disproportionate share of the field's most highly cited outputs, but the reported associations reflect a broad distributional pattern rather than the influence of a small number of exceptional papers. The observed difference nevertheless represents an association between OA status and citation rate rather than evidence of a causal OA effect.

Variation in the Open-Access Citation Association Across Topics

The aggregate comparison does not establish whether the observed association is distributed uniformly across LIS. The OA versus non-OA comparison was therefore repeated within each topic. Of the 52 substantive topics, 50 met the predefined criterion for adequate representation of both OA and non-OA publications and were included in the topic-stratified analysis. Benjamini–Hochberg false-discovery-rate correction [44] was applied across the 50 topic-level tests.

After correction, 10 topics showed a statistically significant positive OA-citation association, while 40 topics showed no statistically detectable association. No topic showed a statistically significant negative association after correction. Table 3 summarizes the topics that remained significant after correction, and Figure 6 places these results in the context of the broader distribution of topic-level estimates.

Table 3 Topics Showing Statistically Significant Positive OA-Citation Associations After FDR Correction

Topic	r	p **(FDR)**	**Significance**
Health Sciences & Clinical Librarianship	.475	$< .001$	Highly sig.
Nursing/Medical Student Info. Literacy	.303	.003	Very sig.
Accessibility & Disability Services	.293	.025	Significant
Academic Library Strategic Planning	.265	.003	Very sig.
Librarian Job Satisfaction & Motivation	.257	.049	Significant
Digital Libraries & Preservation	.236	.005	Very sig.
LIS Education & Practice (Developing Regions)	.228	.012	Significant
Information Retrieval Systems	.195	.017	Significant
Information Seeking Behavior (General)	.167	.008	Very sig.
Information Literacy Instruction	.103	.002	Very sig.

Highly sig.: $p < .001$; Very sig.: $p < .01$; Significant: $p < .05$ (Benjamini-Hochberg FDR-corrected)

The strongest corrected association occurred in Health Sciences & Clinical Librarianship ($r = .475$), followed by Nursing/Medical Student Information Literacy ($r = .303$) and Accessibility & Disability Services ($r = .293$). The remaining significant topics showed smaller positive associations, ranging from $r = .103$ to $r = .265$. The significant topics were concentrated mainly in health-sciences-adjacent and other specialized areas of LIS rather than being distributed evenly across the discipline.

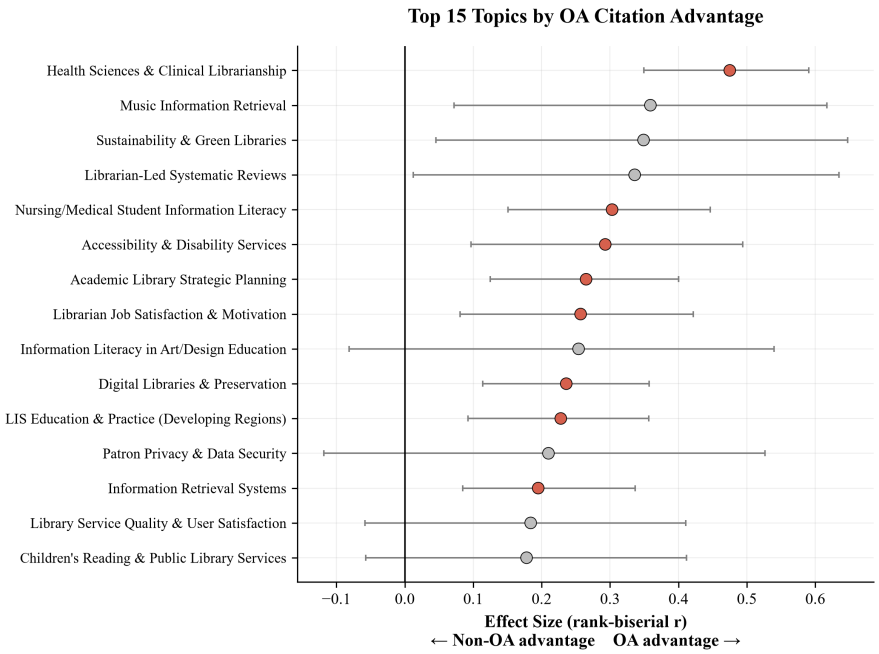


Fig. 6 Open-access citation associations for the fifteen topics with the largest OA associations. Red points indicate significance after FDR correction.

The forest plot also shows that statistical significance and effect magnitude were not equivalent. Several topics had relatively large positive point estimates but did not survive FDR correction, generally accompanied by wider confidence intervals associated with smaller topic-specific samples. The absence of statistically significant negative associations further indicates that the topic-level variation primarily concerns the presence and magnitude of positive associations rather than evidence of a citation disadvantage for OA publications.

A two-way analysis of variance using log-transformed citation rates provided an omnibus assessment of topic heterogeneity. The OA-by-topic interaction was statistically significant, $F(51, 12461) = 1.47$, $p = .017$, indicating that the relationship between OA status and citation rate differed across topics.

Journal-Level Citation Environment and Regression Results

The topic-level analysis established that the OA-citation association varies across LIS research areas.The aggregate association was next examined for sensitivity to publication-level and journal-level characteristics. A sequence of progressively controlled regression models was estimated, beginning with a bivariate OLS model of the OA indicator alone and subsequently adding publication year, author count, international collaboration, document type, and the journal-level citation baseline.

The OA coefficient was 0.149 in the bivariate OLS model, declined to 0.135 after adjustment for publication year, and to 0.114 after adding author count and international collaboration. After document type and the journal-level citation baseline were introduced, the coefficient declined further to 0.054, estimated on a reduced sample of 11,094 publications following exclusion of low-volume journals from the journal-baseline construction (Methods). The OA coefficient remained positive and statistically significant in the fully specified model. The corresponding regression estimates are reported in Table 4.

Table 4 Progressive Regression Models: OA Coefficient and Model Fit

Model	β	**95% CI**	R^2	N
Bivariate (OLS)	0.149	[0.124, 0.173]	.012	12,565
+ Year	0.135	[0.110, 0.160]	.029	12,565
+ Author count, intl. collab.	0.114	[0.090, 0.138]	.060	12,565
+ Document type, journal baseline	0.054	[0.031, 0.076]	.310	11,094

The journal-level citation baseline was the strongest predictor in the fully specified model ($\beta = .235$, $p < .001$). Its introduction produced the largest change in the OA coefficient, reducing the estimate from .114 to .054. In comparison, the addition of publication year, author count, and international collaboration produced smaller changes in the OA coefficient. Figure 7 presents the fully specified model alongside the change in the OA coefficient across the progressive specifications.

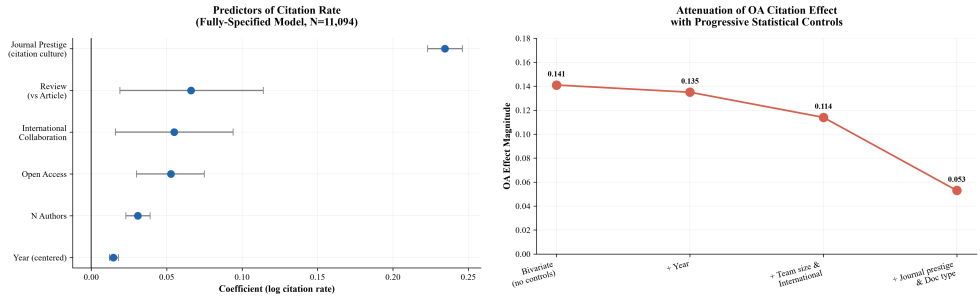


Fig. 7 Regression estimates and attenuation of the open-access citation association. The left panel shows coefficients from the fully specified model and the right panel shows the OA coefficient across progressively controlled models.

The fully specified model was estimated on 11,094 publications. In addition to the journal-level citation baseline, international collaboration, review status, and author count showed positive associations with citation rate. The journal-level baseline had the largest coefficient among the predictors considered. The persistence of a positive OA coefficient after adjustment indicates that the measured journal-level citation environment does not account for the entire observed OA-citation association.

The magnitude of the OA coefficient nevertheless changed substantially across the model sequence. The reduction from .114 to .054 following inclusion of the journal-level baseline indicates that the estimated association is sensitive to the citation environment of the publishing journal. This result does not establish that the journal environment causes the OA association, but it demonstrates that estimates of the association differ materially depending on whether this within-corpus journal-level characteristic is incorporated into the model.

Robustness and Sensitivity Analyses

The principal findings were examined under alternative citation-count specifications and alternative treatment of topic-model outliers. First, a negative binomial model of raw citation counts using log citation age as an exposure offset yielded a positive OA coefficient ($\beta = .162$, $p < .001$), corresponding to an estimated 17.5% higher citation count for OA publications after accounting for exposure time and the other covariates. A separate analysis restricted to papers aged four to six years at the time of analysis ($n = 3,093$) produced an OA coefficient of .057 ($p = .018$), closely comparable to the fully specified primary estimate of .054.

The topic-stratified analysis was also repeated after excluding rather than reassigning documents initially classified by HDBSCAN as unclassified. Six topics remained statistically significant after FDR correction, including Health Sciences & Clinical Librarianship. All six were also significant in the principal analysis, although four additional topics that were significant after outlier reassignment did not remain significant when those documents were excluded. The strongest topic-level association remained positive and of comparable magnitude under the alternative procedure.

The topic-model parameter-selection and stability analyses are reported in the Supplementary Material. These analyses produced broadly consistent topic structures across embedding architectures and bootstrap samples, with greater instability among smaller topics.

The journal-level citation baseline was further tested against two more conservative approaches to journal clustering. Replacing the continuous baseline with a full set of journal fixed effects yielded an OA coefficient of 0.057 ($p < .001$, 95% CI [0.023, 0.091]), with model fit improving substantially ($R^2 = .371$). A multilevel specification with random intercepts for journal yielded a comparable OA coefficient of 0.058 ($p < .001$, 95% CI [0.029, 0.086]). Both estimates closely matched the continuous baseline model's coefficient of 0.054, indicating that the reported association is not an artifact of how journal-level clustering was specified.

A further check tested whether the journal-level baseline's explanatory power was itself confounded with topic, given that certain journals may specialize in specific research topics. Adding topic fixed effects to the fully specified model left the journal-baseline coefficient essentially unchanged (0.234 versus 0.235 without topic controls, $p < .001$), while the OA coefficient declined only modestly, to 0.048 ($p < .001$, 95% CI [0.025, 0.070]), and model fit improved further ($R^2 = .359$). This indicates that the journal-level citation baseline captures a genuine venue-level citation environment distinct from, rather than a proxy for, the discipline's topic structure. Table 5 summarizes the OA coefficient across all four journal-related specifications.

Table 5 The OA Coefficient Under Alternative Journal-Level Specifications

Specification	β	**95% CI**	R^2
Continuous journal baseline (main)	0.054	[0.031, 0.076]	.310
Journal fixed effects	0.057	[0.023, 0.091]	.371
Multilevel (random intercepts)	0.058	[0.029, 0.086]	—
Baseline + topic fixed effects	0.048	[0.025, 0.070]	.359

Temporal Stability

Finally, the magnitude of the OA-citation association was examined for systematic change over the citation-analysis period. The OA-by-year interaction was positive but did not reach conventional statistical significance ($\beta = .0056$, $p = .071$). Thus, the analysis provides no statistically detectable evidence of a systematic change in the OA-citation association between 2010 and 2022. The corresponding temporal visualization is provided in the Supplementary Material.

Discussion

This study examined whether the open access citation association reported at the discipline level holds uniformly within a single field or instead conceals substantial internal variation. The evidence indicates meaningful topic-level differences in the OA-citation association. Eighty percent of the topics examined showed no statistically detectable association between OA status and citation rate, while the overall association was largely accounted for by a smaller subset of topics concentrated in health-sciences-adjacent subfields. The discipline-wide estimate therefore provides a useful overall summary, but it does not fully capture the variation observed across LIS research topics. This finding suggests that interpreting OACA at the discipline level alone may obscure differences in how the association manifests across subfields.

The journal-level baseline and what it means for citation-based evaluation

A second contribution of this study is methodological. The OA coefficient declined from 0.149 in the unadjusted model to 0.054 in the fully specified model, a reduction of more than 60%, with most of the attenuation occurring when the journal-level citation baseline was introduced. This indicates that journal-level citation environment accounts for a substantial portion of the association observed in the unadjusted comparison and highlights the importance of accounting for venue-level differences in OACA research. Studies reporting larger uncontrolled associations, such as the 18% average advantage reported in a multi-million-article analysis [45], may therefore capture differences in citation environment alongside differences associated with OA status. The findings support treating journal-level citation environment as an important consideration in future OACA analyses.

The fully specified model estimates the OA association conditional on the measured covariates and the journal-level citation baseline. It should not, however, be interpreted as a strictly within-journal or venue-independent effect. The baseline is a continuous measure of typical citation performance within the corpus rather than a journal fixed effect, and therefore does not eliminate all sources of journal-level variation. The coefficient of 0.054 should consequently be interpreted as a conditional statistical association rather than as a true venue-independent or causal effect.

This finding also contributes to the broader discussion of citation-based research evaluation. The San Francisco Declaration on Research Assessment [46] cautions against using journal-level metrics as proxies for the quality of individual articles. The present findings provide an empirical illustration of why such caution matters: journal-level citation environment explains substantial variation in citation performance and materially attenuates the observed OA association. Nevertheless, the OA association remains statistically detectable across the specifications examined, although its magnitude is modest.

Why health-sciences-adjacent topics: audience composition and invisible colleges

The significant associations were concentrated in topics such as health sciences and clinical librarianship, nursing and medical student information literacy, and accessibility and disability services. This is consistent with the broader OACA literature, which has generally reported stronger open access associations in STEM-related fields than in the social sciences and humanities. This study extends that observation by showing a similar pattern within a single social-science discipline, at the level of individual research topics.

This pattern was also tested statistically rather than inferred from the significant topics alone. Six topics were classified as health-adjacent using a fixed set of keywords, without considering their statistical results. Three of these six topics showed a significant OA association (50%), compared with seven of the remaining 44 topics (16%). The resulting odds ratio was 5.3, although the difference was not statistically significant in Fisher's exact test ($p = .086$). The result therefore provides suggestive but not conclusive evidence that health-adjacent topics may be more likely to show an OA association. The small number of health-adjacent topics also limits the statistical power of this comparison.

One possible explanation relates to the **invisible college**, a network of researchers and practitioners who read and cite one another's work around a shared area of interest. Health-sciences-adjacent LIS topics may reach both the LIS research community and health professionals such as clinical librarians and allied health practitioners. Some of these readers may have less access to subscription-based LIS journals. Open access could therefore matter more for topics that reach these readers because removing the paywall may expand the group of people able to read and cite the work. This interpretation is consistent with earlier visibility-based explanations of OACA [47, 48].

However, this study did not measure who read these papers, how they accessed them, or whether practitioners were responsible for the additional citations. The invisible-college explanation should therefore be treated as a plausible interpretation

rather than a demonstrated mechanism. Other explanations are also possible, including differences in funding and OA mandates, citing practices in health-related research, article characteristics, or reference patterns. Future research using downloads, altmetrics, or other measures of readership could test whether differences in audience composition help explain the topic-level pattern.

This also explains why the absence of significant associations in topics such as public library value and pandemic-related library services is important. High publication volume alone does not appear to produce an OA association. The results instead suggest that the relationship between OA and citations may depend partly on the audience a research topic reaches.

Topic size, topic sensitivity, and bibliometric concentration

The divergence between the full topic-level analysis and the analysis restricted to the ten highest-volume topics is informative in its own right. The topics driving the overall interaction are, with one exception, not among the discipline's largest or most prolific, but moderate-sized, subject-specific clusters. This is not an isolated curiosity; it is a topic-level analogue of a long-established family of regularities in bibliometrics describing concentration within a minority of a population, most notably Bradford's law of scatter [49], which describes how a small core of sources accounts for a disproportionate share of a field's output, and Lotka's law [50], which describes a comparably skewed distribution of author productivity. The present finding suggests that citation-relevant properties such as the open access association examined here can themselves be distributed according to a similarly skewed logic, concentrated in a minority of a discipline's substantive themes rather than spread in proportion to publication volume. Discipline-wide OACA estimates, the dominant approach in the existing literature, may therefore systematically underweight precisely the subfields where an association is strongest, simply because those subfields do not dominate publication volume and are consequently swamped in an aggregate average by the discipline's larger, less sensitive topics. Several additional topics not reaching significance after correction showed effect sizes of comparable magnitude to the topics that did, with confidence intervals too wide, owing to small subgroup sizes, to reach significance; the true extent of topic-level concentration versus topic-level statistical power limitations cannot be fully disentangled with the present sample and would benefit from replication on a larger or multi-database corpus.

Temporal stability and the shelf life of an access-based citation signal

The absence of a significant change in the OA association across the study period is a modest, supporting result rather than a central finding of this study, but it carries a forward-looking implication worth stating explicitly. Given the substantial growth in open access publishing and policy adoption over this window, a reasonable expectation might have been that any marginal citation association tied to open access status would shrink over time, as openness transitions from a distinguishing signal to a baseline expectation shared by most papers in a field. That no such shift is evident

here suggests either that open access continues to carry a comparable discoverability association even as it becomes more common, or that the present study window has not yet reached the point of saturation at which such a decline would emerge. If the latter proves correct as open access approaches near-universal adoption in the years following this study's window, the open access association itself may have a natural shelf life as a usable citation predictor: a signal that distinguishes a minority of accessible papers from a majority of inaccessible ones loses its distinguishing power once nearly all papers share the same status, an implication with consequences for how long citation-based justifications for open access mandates remain empirically supportable in their current form. The data available here cannot determine which of these possibilities holds, and a longer time series extending further into the present decade would be needed to do so.

Implications

Three implications follow with reasonable directness from these findings. First, for the evaluation of individual authors and papers, a citation advantage attributed to open access status should be treated as topic-dependent and journal-conditional rather than general; on this evidence, the citation-based case for open access is considerably stronger for health-adjacent, practice-facing research than for the discipline's larger, more institutionally central topics. Second, for the design of future OACA research specifically, the size of the attenuation documented here, more than sixty percent of the naive coefficient, indicates that a journal-level baseline is not a minor robustness addition but a load-bearing element of a credible estimate, and topic-stratified analysis is similarly necessary to avoid mistaking a concentrated effect for a general one. Third, for policy that relies on citation-impact arguments to justify open access mandates, including current frameworks that apply uniformly across a funded discipline's constituent topics, this evidence recommends more topic- and field-specific framing than a single, discipline-wide claim can support; institutional experience with transformative agreements illustrates a related but distinct friction, in which formal open-access compliance can rise rapidly while author engagement with associated licensing and reporting requirements lags behind [51], underscoring that policy adoption and behavioral uptake are not the same thing. This is offered as an observation drawn from an associational analysis, not as a judgment on the broader case for open access mandates on equity or public-good grounds, which lies outside the scope of this study.

Limitations

Several limitations should be acknowledged despite the robustness checks reported in the Results. First, this study is observational, and the reported associations should not be interpreted as causal relationships. Authors may be more likely to make papers with greater expected citation potential openly accessible, while unobserved characteristics such as author seniority, institutional reputation, or research funding may be associated with both open access status and citation performance. Consequently, the present study design cannot fully disentangle these potential influences.

Second, funding status was not available in the dataset used in this study. Given the documented relationships between funding, open access status, and citation performance, the absence of funding information may leave some relevant variation unaccounted for in the fully specified model.

Third, open access status was operationalized as a single binary category. Different open access routes may have different relationships with citation performance, while cost-related differences may also affect researchers' ability to publish through fee-based open access routes. Therefore, treating open access as a binary variable may obscure heterogeneity among OA routes and limit the precision of the observed associations[52].

Fourth, the corpus was restricted to Scopus-indexed, English-language literature that met the search criteria specified in the Methods. This restriction may limit the generalizability of the findings to the broader LIS literature, particularly research published in other languages or outside Scopus-indexed sources.

Finally, the proposed invisible-college explanation for the concentration of significant associations in some topics should be regarded as an interpretation consistent with the empirical pattern rather than as a directly tested mechanism. Some topic-level comparisons are also limited by the available statistical power, particularly the comparison involving health-adjacent topics, and should therefore be interpreted cautiously. These limitations mean that the findings should be understood as evidence of statistical associations within the study population and design, rather than as evidence of causal effects or patterns that necessarily generalize to all LIS research.

Conclusion

This study examined whether the widely reported open access citation association operates uniformly within a single scholarly discipline, using Library and Information Science (LIS) as an empirical case. By combining data-driven topic modeling with citation regression adjusted for journal-level citation environment, this study found that the relationship between open access and citation performance is more heterogeneous than discipline-wide estimates suggest. Overall, open access was associated with a modest increase in citation rates in LIS, but this association was concentrated in a minority of research topics, including several health-sciences-adjacent areas. For most topics, no statistically detectable association was observed. These null findings should nevertheless be interpreted cautiously because some topic-level comparisons had limited statistical power.

A key contribution of this study is the finding that accounting for journal-level citation environment materially changes the estimated OA association. The journal-level citation baseline had the largest unstandardized coefficient among the predictors in the fully specified model, and its inclusion reduced the OA coefficient by more than half. This indicates that differences in the citation environments of publishing journals account for a substantial part of the association observed in less adjusted comparisons. Discipline-wide estimates may therefore provide a useful overall summary while still concealing important variation associated with research topic and publication venue.

Taken together, the findings suggest that the OA-citation association should not be treated as a uniform feature of an entire scholarly discipline. Its magnitude and

statistical detectability vary across research topics, while the citation environment of the publishing journal also materially influences the observed association. This study therefore contributes both substantive and methodological evidence for a more nuanced assessment of the relationship between open access and scholarly impact. Future research should incorporate topic-level heterogeneity, journal-level citation environment, funding information, and differentiated OA routes to further examine the factors associated with citation differences. Such work can strengthen the evidence base for OA policy by clarifying when and where open access is associated with greater scholarly visibility and citation impact.

Declarations

- **Funding.** This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

- **Conflict of interest/Competing interests.** The authors declare that they have no known competing financial interests or personal relationships that could have influenced the work reported in this paper.
- **Ethics approval and consent to participate.** This study relied exclusively on publicly available, previously published bibliographic and citation metadata retrieved from Scopus. No human participants, animal subjects, or personal or identifiable data were involved, and no ethics approval was required.
- **Consent for publication.** Not applicable.
- **Data availability.** The primary bibliometric and citation data supporting the findings of this study, including the complete topic structure, topic-level statistical results, and regression outputs, are provided in the Appendices. The full corpus-level dataset is available from the corresponding author upon reasonable request.
- **Materials availability.** Not applicable.
- **Code availability.** The analysis code used to generate the results reported in this study is available from the corresponding author upon reasonable request.
- **Author contribution.** K.S. conceptualized the study, conducted the analysis, developed the methodology, and prepared the main draft. A.K. conducted the literature review and wrote the manuscript. M.S. and D.G. supervised the project and contributed to copyediting and critical revision of the manuscript. All authors read and approved the final manuscript.

References

[1] Dash, S.: The evolution of open access publishing: A longitudinal bibliometric perspective (2004–2024). Serials Librarian **87**(1-2), 31–58 (2026) https://doi.org/10.1080/0361526X.2025.2596290

[2] Moed, H.F.: The effect of "open access" on citation impact: An analysis of arxiv's condensed matter section. Journal of the American Society for Information Science and Technology **58**(13), 2047–2054 (2007) https://doi.org/10.1002/asi.20663

[3] Davis, P.M., Fromerth, M.J.: Does the arxiv lead to higher citations and reduced publisher downloads for mathematics articles? Scientometrics **71**(2), 203–215 (2007) https://doi.org/10.1007/s11192-007-1661-8

[4] Gargouri, Y., Hajjem, C., Larivière, V., Gingras, Y., Carr, L., Brody, T., Harnad, S.: Self-selected or mandated, open access increases citation impact for higher quality research. PLoS ONE **5**(10) (2010) https://doi.org/10.1371/journal.pone.0013636

[5] Craig, I.D., Plume, A.M., McVeigh, M.E., Pringle, J., Amin, M.: Do open access articles have greater citation impact? a critical review of the literature. Journal of Informetrics **1**(3), 239–248 (2007) https://doi.org/10.1016/j.joi.2007.04.001

[6] Langham-Putrow, A., Bakker, C., Riegelman, A.: Is the open access citation advantage real? a systematic review of the citation of open access and subscription-based articles. PLOS ONE **16**(6) (2021) https://doi.org/10.1371/journal.pone.0253129

[7] Huang, C.-K., Neylon, C., Montgomery, L., Hosking, R., Diprose, J.P., Handcock, R.N., Wilson, K.: Open access research outputs receive more diverse citations. Scientometrics **129**(2), 825–845 (2024) https://doi.org/10.1007/s11192-023-04894-0

[8] Miguel, S., Chinchilla-Rodríguez, Z., De Moya-Añegón, F.: Open access and scopus: A new approach to scientific visibility from the standpoint of access. Journal of the American Society for Information Science and Technology **62**(6), 1130–1145 (2011) https://doi.org/10.1002/asi.21532

[9] Warlick, S.E., Vaughan, K.T.L.: Factors influencing publication choice: Why faculty choose open access. Biomedical Digital Libraries **4** (2007) https://doi.org/10.1186/1742-5581-4-1

[10] Holmberg, K., Hedman, J., Bowman, T.D., Didegah, F., Laakso, M.: Do articles in open access journals have more frequent altmetric activity than articles in subscription-based journals? an investigation of the research output of finnish universities. Scientometrics **122**(1), 645–659 (2020) https://doi.org/10.1007/s11192-019-03301-x

[11] Ottaviani, J.: The post-embargo open access citation advantage: It exists (probably), it's modest (usually), and the rich get richer (of course). PLoS ONE **11**(8), 0159614 (2016) https://doi.org/10.1371/journal.pone.0159614

[12] Clayson, P.E., Baldwin, S.A., Larson, M.J.: The open access advantage for studies of human electrophysiology: Impact on citations and altmetrics. International Journal of Psychophysiology **164**, 103–111 (2021) https://doi.org/10.1016/j.ijpsycho.2021.03.006

[13] Clark, A.D., Myers, T.C., Steury, T.D., Krzton, A., Yanes, J., Barber, A., Barry,

J., Barua, S., Eaton, K., Gosavi, D., Nance, R., Pervaiz, Z., Ugochukwu, C., Hartman, P., Stevison, L.S.: Does it pay to pay? a comparison of the benefits of open-access publishing across various sub-fields in biology. PeerJ **12** (2024) https://doi.org/10.7717/peerj.16824

[14] Fleming, J.I., Cook, B.G.: Examining a citation advantage for publicly accessible articles in special education journals. Royal Society Open Science **13**(6), 1–13 (2026) https://doi.org/10.1098/rsos.252032

[15] Friederichs, H.: Is it worth publishing open access?—the scientific impact of open access publications in the field of medical education. Medical Education Online **31**(1) (2026) https://doi.org/10.1080/10872981.2026.2652722

[16] Basson, I., Blanckenberg, J.P., Prozesky, H.: Do open access journal articles experience a citation advantage? results and methodological reflections of an application of multiple measures to an analysis by wos subject areas. Scientometrics **126**(1), 459–484 (2021) https://doi.org/10.1007/s11192-020-03734-9

[17] Lin, W.-Y.C.: Effects of open access and articles-in-press mechanisms on publishing lag and first-citation speed: a case on energy and fuels journals. Scientometrics **126**(6), 4841–4869 (2021) https://doi.org/10.1007/s11192-021-03956-5

[18] Khan, D., Ashar, M., Yuvaraj, M.: Do open access journals have a greater citation impact? a study of journals in library and information science. Collection and Curation **42**(1), 13–24 (2023) https://doi.org/10.1108/CC-03-2022-0010

[19] Lee, S., Shim, W.: An empirical analysis of open access citation advantages in library and information science. Information Research **31**(iConf), 1733–1750 (2026) https://doi.org/10.47989/ir31iConf64133

[20] Dorta-González, P., Dorta-González, M.I.: Citation differences across research funding and access modalities. Journal of Academic Librarianship **49**(4) (2023) https://doi.org/10.1016/j.acalib.2023.102734

[21] Momeni, F., Dietze, S., Mayr, P., Biesenbender, K., Peters, I.: Which factors are associated with open access publishing? a springer nature case study. Quantitative Science Studies **4**(2), 353–371 (2023) https://doi.org/10.1162/qss_a_00253

[22] Borrego, Á.: Article processing charges for open access journal publishing: A review. Learned Publishing **36**(3), 359–378 (2023) https://doi.org/10.1002/leap.1558

[23] Maddi, A., Sapinho, D.: On the culture of open access: the sci-hub paradox. Scientometrics **128**(10), 5647–5658 (2023) https://doi.org/10.1007/s11192-023-04792-5

[24] Perry, G.T., Pereira, L.S.: Do open-access articles have a citation advantage?

a comparative analysis on hybrid journals pre and post the cessation of sci-hub updates. Scientometrics **131**(5), 3169–3191 (2026) https://doi.org/10.1007/s11192-026-05625-x

[25] Colavizza, G., Cadwallader, L., LaFlamme, M., Dozot, G., Lecorney, S., Rappo, D., Hrynaszkiewicz, I.: An analysis of the effects of sharing research data, code, and preprints on citations. PLoS ONE **19**(10) (2024) https://doi.org/10.1371/journal.pone.0311493

[26] Pan, X., Li, K., Kuang, Q., Ma, Z.: Does open data promote the impact of research articles in psychology? Journal of Information Science (2026) https://doi.org/10.1177/01655515261461729

[27] Wijewickrema, M.: A bibliometric study on library and information science and information systems literature during 2010–2019. Library Hi Tech **41**(2), 595–621 (2023) https://doi.org/10.1108/LHT-06-2021-0198

[28] Siddique, N., Rehman, S.U., Khan, M.A., Altaf, A.: Library and information science research in pakistan: A bibliometric analysis, 1957–2018. Journal of Librarianship and Information Science **53**(1), 89–102 (2021) https://doi.org/10.1177/0961000620921930

[29] Siddique, N., Ur Rehman, S., Ahmad, S., Abbas, A., Khan, M.A.: Library and information science research in the arab world: a bibliometric analysis 1951–2021. Global Knowledge, Memory and Communication **72**(1-2), 138–159 (2023) https://doi.org/10.1108/GKMC-06-2021-0103

[30] Mokhtari, H., Barkhan, S., Haseli, D., Saberi, M.K.: A bibliometric analysis and visualization of the journal of documentation: 1945–2018. Journal of Documentation **77**(1), 69–92 (2021) https://doi.org/10.1108/JD-08-2019-0165

[31] Hussain, A., Ahmad, S.: Mapping the literature on artificial intelligence in academic libraries: A bibliometrics approach. Science and Technology Libraries **43**(2), 131–146 (2024) https://doi.org/10.1080/0194262X.2023.2238198

[32] Gao, Q., Huang, X., Dong, K., Liang, Z., Wu, J.: Semantic-enhanced topic evolution analysis: a combination of the dynamic topic model and word2vec. Scientometrics **127**(3), 1543–1563 (2022) https://doi.org/10.1007/s11192-022-04275-z

[33] Wang, Z., Chen, J., Chen, J., Chen, H.: Identifying interdisciplinary topics and their evolution based on bertopic. Scientometrics **129**(11), 7359–7384 (2024) https://doi.org/10.1007/s11192-023-04776-5

[34] Mustak, M., Salminen, J., Plé, L., Wirtz, J.: Artificial intelligence in marketing: Topic modeling, scientometric analysis, and research agenda. Journal of Business Research **124**, 389–404 (2021) https://doi.org/10.1016/j.jbusres.2020.10.044

[35] Lee, C.-H., Liu, C.-L., Trappey, A.J.C., Mo, J.P.T., Desouza, K.C.: Understanding digital transformation in advanced manufacturing and engineering: A bibliometric analysis, topic modeling and research trend discovery. Advanced Engineering Informatics **50** (2021) https://doi.org/10.1016/j.aei.2021.101428

[36] Mariani, M., Baggio, R.: Big data and analytics in hospitality and tourism: a systematic literature review. International Journal of Contemporary Hospitality Management **34**(1), 231–278 (2022) https://doi.org/10.1108/IJCHM-03-2021-0301

[37] Chen, X., Zou, D., Xie, H., Cheng, G.: Twenty years of personalized language learning: Topic modeling and knowledge mapping. Educational Technology and Society **24**(1), 205–222 (2021)

[38] Chen, X., Zou, D., Xie, H., Cheng, G., Liu, C.: Two decades of artificial intelligence in education: Contributors, collaborations, research topics, challenges, and future directions. Educational Technology and Society **25**(1), 28–47 (2022)

[39] Chen, X., Zou, D., Xie, H., Wang, F.L.: Past, present, and future of smart learning: a topic-based bibliometric analysis. International Journal of Educational Technology in Higher Education **18**(1) (2021) https://doi.org/10.1186/s41239-020-00239-6

[40] Chen, X., Zou, D., Xie, H., Wang, F.L.: Metaverse in education: Contributors, cooperations, and research themes. IEEE Transactions on Learning Technologies **16**(6), 1111–1129 (2023) https://doi.org/10.1109/TLT.2023.3277952

[41] Velez-Estevez, A., García-Sánchez, P., Moral-Munoz, J.A., Cobo, M.J.: Why do papers from international collaborations get more citations? a bibliometric analysis of library and information science papers. Scientometrics **127**(12), 7517–7555 (2022) https://doi.org/10.1007/s11192-022-04486-4

[42] Macháček, V., Srholec, M.: Predatory publishing in scopus: Evidence on cross-country differences. Quantitative Science Studies **3**(3), 859–887 (2022) https://doi.org/10.1162/qss_a_00213

[43] Grootendorst, M.: Bertopic: Neural topic modeling with a class-based tf-idf procedure. arXiv preprint arXiv:2203.05794 (2022) https://doi.org/10.48550/arXiv.2203.05794

[44] Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society: Series B (Methodological) **57**(1), 289–300 (1995) https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

[45] Piwowar, H., Priem, J., Larivière, V., Alperin, J.P., Matthias, L., Norlander, B., Farley, A., West, J., Haustein, S.: The state of oa: a large-scale analysis of

the prevalence and impact of open access articles. PeerJ **6**, 4375 (2018) https://doi.org/10.7717/peerj.4375

[46] Cagan, R.: The san francisco declaration on research assessment. Disease Models and Mechanisms **6**(4), 869–870 (2013) https://doi.org/10.1242/dmm.012955

[47] Antelman, K.: Do open-access articles have a greater research impact? College and Research Libraries **65**(5), 372–382 (2004) https://doi.org/10.5860/crl.65.5.372

[48] Wang, X., Liu, C., Mao, W., Fang, Z.: The open access advantage considering citation, article usage and social media attention. Scientometrics **103**(2), 555–564 (2015) https://doi.org/10.1007/s11192-015-1547-0

[49] Bradford, S.C.: Sources of information on specific subjects. Engineering **137**, 85–86 (1934). Reprinted in *Journal of Information Science*, 10(4), 173–175 (1985), doi:10.1177/016555158501000406

[50] Lotka, A.J.: The frequency distribution of scientific productivity. Journal of the Washington Academy of Sciences **16**(12), 317–323 (1926)

[51] Guo, Y., Dou, T., Zhang, S., Zhang, C., Li, Q., Jin, J.: Monitoring the implementation of the acm-thu open transformative agreement: Institutional practices and insights from tsinghua university. Lecture Notes in Computer Science **16242**, 309–323 (2026) https://doi.org/10.1007/978-981-95-4861-3_26

[52] Williams, J.W., Taylor, A., Tolley, K.A., Provete, D.B., Correia, R., Guedes, T.B., Farooq, H., Li, Q., Pinheiro, H.T., Liz, A.V., Luna, L.W., Matthews, T.J., Palmeirim, A.F., Puglielli, G., Rivadeneira, M.M., Robin, V.V., Schrader, J., Shestakova, T.A., Tukiainen, H., Heyden, S., Zizka, A.: Shifts to open access with high article processing charges hinder research equity and careers. Journal of Biogeography **50**(9), 1485–1489 (2023) https://doi.org/10.1111/jbi.14697

Appendix A Complete Topic Structure (All 52 Topics)

Table A.1 presents the complete set of 52 substantive research topics identified by BERTopic, extending Table 1 in the main text to the full topic structure. Document counts, corpus shares, and representative keywords are drawn directly from the topic model's class-based TF-IDF output.

Table A.1: Complete List of 52 Research Topics with Document Counts and Representative Keywords

Rank	Research topic	n	%	Representative terms
1	Information Literacy Instruction	2,344	15.92	information literacy; literacy; information; students; il; learning; instruction; skills; teaching; school
2	Public Libraries - Community & Value	977	6.64	public; public libraries; public library; libraries; library; social; community; services; value; study
3	LIS Education & Practice (Developing Regions)	855	5.81	library; study; nigeria; libraries; ict; university; librarians; resources; services; lis
4	Information Seeking Behavior (General)	810	5.50	seeking; information; information seeking; behaviour; behavior; information behaviour; study; research; seeking behaviour; information behavior
5	Health Information Seeking Behavior	791	5.37	health; seeking; information seeking; health information; information; cancer; risk; 19; covid; covid 19
6	Academic Librarians & Faculty Status	749	5.09	librarians; academic; research; faculty; librarian; academic librarians; professional; library; librarianship; job
7	LIS Bibliometrics & Scholarly Communication	736	5.00	lis; science; journals; information science; research; library information; articles; journal; citation; authors
8	Library Collections & E-books	482	3.27	books; collection; book; academic; libraries; academic libraries; collections; library; print; electronic
9	Information Retrieval Systems	415	2.82	retrieval; information retrieval; ir; query; search; based; systems; information; documents; semantic
10	Digital Libraries & Preservation	392	2.66	digital; digital library; digital libraries; library; libraries; preservation; collections; archives; metadata; heritage
11	Librarian Job Satisfaction & Motivation	336	2.28	job; librarians; job satisfaction; knowledge; satisfaction; performance; work; study; management; stress
12	Academic Library Strategic Planning	330	2.24	academic; academic libraries; libraries; change; strategic; library; planning; academic library; management; future
13	Embedded/Distance Learning Librarianship	326	2.21	embedded; distance; online; learning; embedded librarianship; librarian; librarians; students; librarianship; embedded librarian

Table A.1 – continued

Rank	Research topic	n	%	Representative terms
14	Health Sciences & Clinical Librarianship	307	2.09	health; medical; health sciences; sciences; clinical; librarians; hospital; librarian; care; sciences librarians
15	COVID-19 & Library Services	296	2.01	covid; covid 19; 19; pandemic; 19 pandemic; services; libraries; library; online; academic
16	Social Media Use in Libraries	264	1.79	social media; social; media; facebook; use social; use; libraries; twitter; networking; social networking
17	Research Data Management	249	1.69	data; research data; data management; research; big data; big; management; services; libraries; academic
18	Nursing/Medical Student Information Literacy	218	1.48	nursing; students; medical; evidence; information literacy; literacy; information; evidence based; skills; nursing students
19	Theoretical Foundations of Information Science	204	1.39	science; information science; information; knowledge; library information; field; theory; research; theoretical; concept
20	Accessibility & Disability Services	201	1.37	disabilities; accessibility; impaired; visually; disability; library; services; blind; disabled; users
21	Library Service Quality & User Satisfaction	200	1.36	service; quality; service quality; satisfaction; library; services; users; customer; loyalty; user
22	Open Access & Scholarly Publishing	199	1.35	open; open access; access; journals; publishing; scholarly; repositories; oer; institutional; academic
23	Children's Reading & Public Library Services	188	1.28	children; reading; public; public library; young; public libraries; early; library; fiction; libraries
24	Academic Library Spaces & Makerspaces	181	1.23	space; spaces; library; academic library; academic; learning; users; students; libraries; makerspaces
25	Diversity, Equity & Inclusion in LIS	179	1.22	diversity; racial; inclusion; justice; dei; social justice; librarians; color; equity; american
26	Library Websites & Usability	173	1.18	web; websites; website; library websites; library; libraries; academic; usability; user; academic libraries

Table A.1 – continued

Rank	Research topic	n	%	Representative terms
27	Mobile Technology in Libraries	164	1.11	mobile; devices; services; mobile devices; library; mobile library; use; library services; mobile technology; apps
28	Consumer/Public Health Literacy	160	1.09	health; health information; health literacy; public; consumer health; consumer; literacy; information; public libraries; libraries
29	Virtual/Reference Desk Services	140	0.95	reference; service; chat; reference services; reference service; virtual reference; reference librarians; desk; questions; virtual
30	Academic Library Leadership	130	0.88	leadership; leaders; academic; library leaders; library; library leadership; transformational; academic libraries; academic library; libraries
31	Public Libraries in India	114	0.77	public; public libraries; libraries; india; public library; government; library; district; services; bengal
32	Academic Libraries & Student Retention	109	0.74	student; retention; academic; success; library; student success; academic library; students; learning; academic libraries
33	Semantic Web & Linked Data in Digital Libraries	104	0.71	discovery; semantic; search; digital; digital libraries; ontology; linked; web; linked data; metadata
34	Music Information Retrieval	104	0.71	music; musical; creativity; information; retrieval; information seeking; seeking; information retrieval; creative; librarianship
35	AI & ChatGPT in Libraries	103	0.70	ai; artificial intelligence; artificial; intelligence; intelligence ai; chatgpt; libraries; academic; technology; services
36	Digital Library Acceptance & Adoption	103	0.70	digital library; digital; acceptance; intention; model; use; technology; factors; adoption; perceived
37	Geographic Information Science (GIS)	101	0.69	gis; geographic; geographic information; spatial; map; data; maps; science; information science; information
38	Science Librarianship (STEM-specific)	99	0.67	prize; science librarians; science; quantum; physics; science science; chemistry; taylor francis; francis; taylor

Table A.1 – continued

Rank	Research topic	n	%	Representative terms
39	Library Management Software (ILMS/Open Source)	83	0.56	open source; software; koha; source; open; library management; management; source software; library; integrated library
40	LGBTQ+ Representation in Library Collections	79	0.54	lgbtq; identity; gender; gay; needs; sexual; public; people; libraries; public libraries
41	Librarian-Led Systematic Reviews	74	0.50	systematic; reviews; systematic reviews; review; systematic review; librarian; search; librarians; literature; searching
42	Library Marketing	72	0.49	marketing; services; library; library services; libraries; library marketing; marketing library; brand; marketing strategies; academic
43	Law Librarianship	71	0.48	law; legal; law librarians; law library; law librarianship; librarians; law libraries; legal information; article; law librarian
44	Sustainability & Green Libraries	66	0.45	sustainable; sustainability; green; environmental; sustainable development; sdgs; development; libraries; development goals; goals
45	Copyright Literacy & Law	61	0.41	copyright; law; copyright law; laws; infringement; copyright issues; fair; fair use; libraries; librarians
46	Gamification in Information Literacy	60	0.41	games; game; learning; gaming; information literacy; literacy; video; students; gamification; information
47	Information Literacy in Art/Design Education	60	0.41	art; visual; creative; society north; art libraries; fashion; america rights; literacy; libraries society; design
48	Political Information Seeking Behavior	58	0.39	political; election; political information; seeking; information seeking; information; online; media; participation; presidential
49	Library Disaster Preparedness	57	0.39	disaster; preparedness; crisis; libraries; disasters; public; disaster preparedness; management; public libraries; community
50	Patron Privacy & Data Security	56	0.38	privacy; protection; security; policies; patron; tracking; patron privacy; libraries; data; party

Table A.1 – continued

Rank	Research topic	n	%	Representative terms
51	Cloud Computing in Libraries	46	0.31	cloud; computing; cloud computing; cloud based; adoption; libraries; based; services; library; wearable
52	Private Information Retrieval (CS)	44	0.30	capacity; private; messages; servers; message; retrieval; information retrieval; problem; ieee; storage

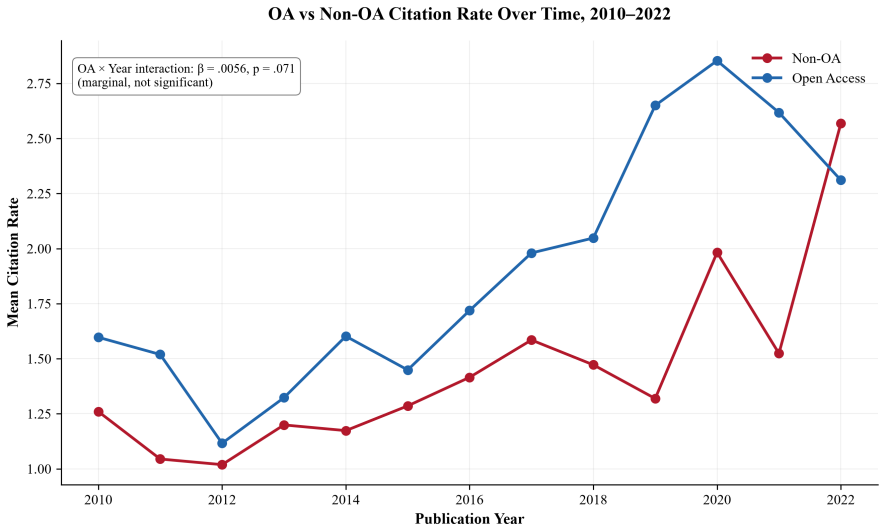


Fig. A.1 Mean citation rate by open-access status across publication years, 2010–2022, corresponding to the temporal stability analysis reported in the Results.

Appendix B Complete Topic-Level Statistical Results (All 50 Tested Topics)

Table B.1 reports the complete Mann-Whitney U topic-stratified results underlying the analysis in the main text, extending Table 3 from the ten statistically significant topics to all 50 topics with adequate subgroup sizes for testing. Effect sizes are signed such that positive values indicate an open access citation advantage, consistent with the convention used in the main text.

Table B.1: Complete Topic-Stratified OA-Citation Association Results

Topic	n_{OA}	n_{non-OA}	**Effect** (r)	**95% CI**	p **(FDR)**	**Sig.**
Health Sciences & Clinical Librarianship	121	153	0.475	[0.350, 0.591]	6.57e-10	Yes
Music Information Retrieval	20	70	0.359	[0.072, 0.617]	0.067	No
Sustainability & Green Libraries	17	31	0.349	[0.045, 0.647]	0.186	No
Librarian-Led Systematic Reviews	34	19	0.336	[0.012, 0.635]	0.186	No
Nursing/Medical Student Information Literacy	91	103	0.303	[0.151, 0.447]	0.003	Yes
Accessibility & Disability Services	42	126	0.293	[0.097, 0.494]	0.025	Yes
Academic Library Strategic Planning	92	202	0.265	[0.125, 0.400]	0.003	Yes
Librarian Job Satisfaction & Motivation	39	250	0.257	[0.081, 0.422]	0.049	Yes
Information Literacy in Art/Design Education	19	35	0.254	[-0.081, 0.540]	0.379	No
Digital Libraries & Preservation	99	254	0.236	[0.114, 0.357]	0.005	Yes
LIS Education & Practice (Developing Regions)	70	670	0.228	[0.093, 0.357]	0.012	Yes
Patron Privacy & Data Security	14	35	0.210	[-0.119, 0.527]	0.573	No
Information Retrieval Systems	115	252	0.195	[0.085, 0.337]	0.017	Yes
Library Service Quality & User Satisfaction	23	145	0.184	[-0.058, 0.411]	0.411	No
Children's Reading & Public Library Services	36	130	0.178	[-0.057, 0.412]	0.341	No
Information Seeking Behavior (General)	173	542	0.167	[0.062, 0.271]	0.008	Yes
Science Librarianship (STEM-specific)	15	74	0.159	[-0.150, 0.497]	0.612	No
Embedded/Distance Learning Librarianship	71	234	0.131	[-0.042, 0.283]	0.338	No
Academic Library Spaces & Makerspaces	59	102	0.129	[-0.062, 0.315]	0.438	No
Semantic Web & Linked Data in Digital Libraries	35	61	0.118	[-0.116, 0.348]	0.612	No
Academic Libraries & Student Retention	24	75	0.114	[-0.152, 0.398]	0.671	No
Copyright Literacy & Law	12	39	0.111	[-0.239, 0.479]	0.819	No

Table B.1 – continued

Topic	n_{OA}	n_{non-OA}	Effect (r)	95% CI	p (FDR)	Sig.
Consumer/Public Health Literacy	47	91	0.108	[-0.082, 0.291]	0.612	No
Political Information Seeking Behavior	13	39	0.103	[-0.245, 0.436]	0.819	No
Information Literacy Instruction	780	1257	0.103	[0.049, 0.155]	0.002	Yes
LIS Bibliometrics & Scholarly Communication	160	449	0.076	[-0.025, 0.175]	0.411	No
Private Information Retrieval (CS)	18	23	0.075	[-0.300, 0.403]	0.904	No
AI & ChatGPT in Libraries	9	22	0.071	[-0.414, 0.596]	0.925	No
Library Websites & Usability	26	134	0.068	[-0.165, 0.302]	0.819	No
Geographic Information Science (GIS)	27	66	0.066	[-0.191, 0.330]	0.838	No
Research Data Management	89	123	0.065	[-0.080, 0.210]	0.672	No
Public Libraries - Community & Value	237	570	0.056	[-0.026, 0.150]	0.491	No
Open Access & Scholarly Publishing	66	100	0.055	[-0.121, 0.234]	0.819	No
Academic Librarians & Faculty Status	229	428	0.054	[-0.042, 0.150]	0.573	No
Library Marketing	6	58	0.049	[-0.368, 0.480]	0.934	No
Health Information Seeking Behavior	200	392	0.048	[-0.041, 0.141]	0.612	No
Library Collections & E-books	103	333	0.046	[-0.086, 0.178]	0.743	No
Gamification in Information Literacy	20	31	0.039	[-0.308, 0.335]	0.934	No
Social Media Use in Libraries	41	196	0.035	[-0.141, 0.231]	0.904	No
Theoretical Foundations of Information Science	68	99	0.027	[-0.156, 0.204]	0.925	No
Library Disaster Preparedness	8	37	0.020	[-0.456, 0.483]	0.941	No
Digital Library Acceptance & Adoption	13	70	0.015	[-0.326, 0.337]	0.941	No
Virtual/Reference Desk Services	42	84	-0.014	[-0.227, 0.190]	0.934	No
Academic Library Leadership	35	80	-0.019	[-0.252, 0.204]	0.934	No
COVID-19 & Library Services	57	147	-0.019	[-0.187, 0.166]	0.934	No

Table B.1 – continued

Topic	n_{OA}	n_{non-OA}	Effect (r)	95% CI	p (FDR)	Sig.
Public Libraries in India	9	90	-0.031	[-0.338, 0.267]	0.934	No
Diversity, Equity & Inclusion in LIS	58	78	-0.037	[-0.237, 0.177]	0.904	No
Mobile Technology in Libraries	40	107	-0.102	[-0.308, 0.118]	0.612	No
LGBTQ+ Representation in Library Collections	18	42	-0.148	[-0.439, 0.159]	0.639	No
Library Management Software (ILMS/Open Source)	15	60	-0.256	[-0.534, 0.036]	0.379	No

Appendix C Journal-Level Citation Baseline: Supplementary Visualization

Figure C.1 provides a decile-binned view of the relationship between the journal-level citation baseline and individual paper citation rates, supporting the regression finding reported in the main text.

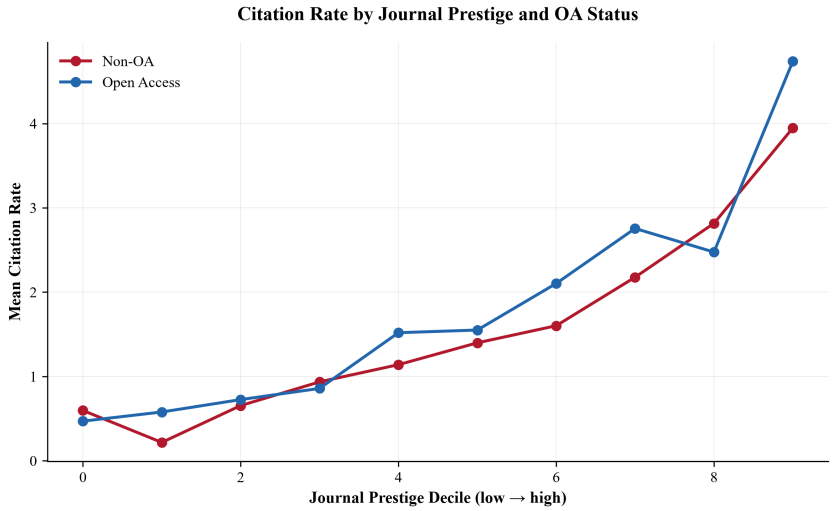


Fig. C.1 Mean citation rate by journal-level citation baseline decile and open-access status.