Electronic Publication Makes Science and Scholarship More Narrow

Summary:

As journals become available electronically, scientists and scholars have more articles at their fingertips, but they cite fewer, more recent ones.

James A. Evans[1]

[1] Department of Sociology, University of Chicago, 1126 E. 59th Street, Chicago, IL 60615.
jevans@uchicago.edu

Online journals promise to serve more information to more dispersed audiences and are more efficiently searched and recalled. But because they are used differently than print—scientists tend to search electronically rather than browse or peruse—electronically available journals portend an ironic change for science. Using a database of 34 million articles, their citations (1945-2005) and online availability (1998-2005), I demonstrate that as more journal issues come online, 1) the articles referenced tend to be more recent, 2) fewer journals and articles are cited and 3) more of those citations are to fewer journals and articles. The forced browsing of print archives may have stretched scholars to anchor findings deeply into past and present scholarship. Searching online is more efficient, and so helps scientists present their findings in a more focused way, but this also increases the specialization of science and scholarship.

Scholarship about “open access,” “digital libraries,” and “information technology” has focused on the superiority of the electronic provision of research. A recent Panel Report from the U.S. President’s Information Technology Advisory Committee (PITAC), *Digital Libraries: Universal Access to Human Knowledge*, captures the tone: “All citizens anywhere anytime can use any Internet-connected digital device to search all of human knowledge….In this vision, no class-room, group, or person is ever isolated from the world’s greatest knowledge resources” (*1*). In recent debate over open access, the message is only slightly more subtle: “Students are well advised to keep doing what they do naturally: favour material that is freely accessible on the Web” (*2*). This overlooks the nature of the interface between the user and the information (*3*); there has been little discussion of browsing/searching technology or its potential effect on science and scholarship.

Recent research into the practice of library usage measures the use of print and electronic resources with surveys, database access logs, circulation records, and reshelving counts. Despite differences in methodology, researchers agree that print use is declining as electronic use increases (*4*), and that general users prefer online material to print (*5*). These studies are also in general agreement about the three most common practices used by scientists and scholars who publish. First, most experts browse or briefly scan a small number of core journals in print or online to build awareness of current research (*6*). After relevant articles are discovered online, these are often printed and perused in depth on paper (*7*). A second practice is to search by topic in an online article database. In recent years, the percentage of papers read as a result of browsing has dropped and been replaced by the results of online searches, especially for the most productive scientists and scholars (*8*). Finally, subject experts use hyperlinks in online articles to view referenced or related articles (*6*). Disciplinary differences exist. For example, biologists

prefer to browse online, while medical professionals place a premium on purchasing and browsing in print. In sum, researchers peruse in print, browse in print or online (*9*), search and follow citations online. These findings follow from the organization and accessibility of print and online papers. Print holdings reside either in a physical "stack" by journal and topic, arranged historically, or in a "recent publications" area. For print journals, the table of contents—its list of titles and authors—serves as primary index. Online archives allow people to browse within journals, but they also facilitate searching the entire archive of available journals. In online interfaces where searching and browsing are both options (e.g., 3 ProQuest, Ovid, EBSCO, JSTOR, etc.), the searching option (e.g., button) is almost always placed first on the interface because logs demonstrate more frequent usage. When searched as an undifferentiated archive of papers, titles, abstracts, and sometimes the full-text can be searched by relevance, by date. Because electronic indexing is richer, experts may still browse in print, but they search online (*10*).

What is the effect of online availability of journal issues? It is possible that by making more research more available, online searching could conceivably broaden the work cited and lead researchers, as a collective, away from the "core" journals of their fields and to dispersed but individually relevant work. I will show, however, that even as deeper journal back-issues become available online, scientists and scholars will cite more recent articles; Even as more total journals become available online, fewer will be cited.

Citation data were drawn from Thompson Scientific's *Science*, Social *Science* and Humanities Citation Indexes, the most complete source of citation data available. Citation Index (CI) data currently includes articles and associated citations from the 6000 most highly cited journals in the sciences, social sciences and humanities going back as far as 1945, for a total of over 50

million articles. The CI flags more than 98% of its journals with from one to three of a possible 300 content codes, such as “condensed matter physics,” “ornithology,” and “inorganic and nuclear chemistry.” Citation patterns were then linked with data tracking the online availability of journals from Information Today, Inc.'s Fulltext Sources Online (FSO).

FSO is the oldest and largest publication about electronic journal availability. Information Today began publishing FSO biannually in 1998, indicating which journals were available in which commercial electronic archives (e.g., Lexis-Nexis, EBSCO, Ovid, etc.) or if available freely on their own website, and for how many back issues. Figure 1 shows the speed of the shift toward commercial and free electronic provision of articles, and how deepening backfiles have made more early science readily available. Merged together by issn, the CI and FSO data allow us to capture how article online availability, by subscription and for free, changes how that knowledge is used in subsequent research. This combined dataset results in 26,002,796 articles whose journals come online by 2006 and a distinct 8,090,813 (in addition to the 26 million) that reference them. The data are ideal for examining the widely discussed influence of open access on citation impact (*11*, *12*), but this will not be ventured here.

Panel regression models were used to explore the relationship between online article availability and citation activity—average historical depth of citations, number of distinct articles and journals cited, and Herfindahl concentration of citations to particular articles and journals—over time. Because studies show substantial variation in reading and research patterns by area, we use fixed effect specifications to compare journals and subfields only to themselves over time as their online availability shifts. In this way, the pattern of citations to a journal or subfield is compared when available only in print, in print and online through a commercial archive, and online for free.

The first question was whether depth of citation—years between articles and the work they reference—is predicted by the depth of journal issues online—how many years back issues were electronically available during the previous year when scientists presumably drafted them into their papers. For subfields, this is calculated as years from the first journal's availability. These data were collected in publication windows of 20 years, and so only data from 1965—20 years after the beginning of the dataset—was used. For the entire dataset, citations pointed to prior articles published an average of 5.6 years previous. The average number of years journal articles were available online is only 1.85—the data goes back to 1945—but with a standard deviation of 5 years and a maximum of over 60 years. Analysis was performed by citation year and within journal or subfield. The standard ordinary-least-squares (OLS) method for linear regression was used in generating all the results to be described. All regression models contain variables used to account and statistically control for alternate explanations of why citations might refer to more recent articles. A sequence of integers from 1 to 40, corresponding to citation years 1965 through 2005, was included to account for a general trend of increasing citations over time (the estimates for this variable were always positive and statistically significant, $p<.001$). Average number of pages and average number of references in citing articles were both included to account for the possibility that citations are more recent because articles are shorter with fewer references and the earliest ones have been disproportionately "censored" by publishers (the estimates for *pages* are positive but not always significant; those for *references* are always positive and significant, $p<.001$: Longer articles with more references refer to earlier work). A measure of the average age of title words was also included in the models to account for the possibility that in recent years, research has concerned more recent concepts or recently discovered (or invented) phenomena. This was calculated by taking the age of each title word

within the relevant publication window for the analysis (e.g., prior 20 years) and then multiplying it by a weight for each word *i* in title *j* such that $\sum_{i=1}^{k}\left(1+\ln\left(tf_{ij}\right)\right)\times\ln\left(\frac{N}{df_i}\right)$ where tf_{ij} equals the frequency of term *i* in title *j* and df_i equals the number of articles in a given year that contain term *i* out of the total number of annual articles *N* (*13*). This approach highly weights distinguishing title terms (e.g., *buckeyballs*, *microRNA*) and gives lesser weight to broad area terms (e.g., *gene*, *ocean*) and virtually no weight to universal words (e.g., *and*, *the*). Regression coefficients for the title age measures are always positive and significant ($p<.0001$), indicating that titles with older terms reference earlier articles. Each model also contains a constant with a significant negative estimate.

Figure 2, panel A illustrates the simultaneous effect of commercial and free online availability on the average age of citations. Consider a journal whose articles reference prior work that is, on average, 5.6 years old—the sample mean. If that journal's issues become available online for an additional 15 years, both commercially and for free, the average age of references would drop to less than 4.5 years, falling by .088 years for each new online year available. The within subfield models follow the same pattern, although confidence intervals are naturally wider. Notable in both models is that free online availability drives citation depth down more than twice as rapidly as commercial availability.

To test for the effect of online availability on the amount of distinct research cited, I explored the relationship between the distinct number of articles and journals cited in a given citation year by depth of online availability. Number of distinct articles and journals was calculated over a 20 year window, following the prior analysis. For the average journal, 632 articles are cited each year, but this ranges widely. Because citation values are discrete and because high values concentrate within a few core journals but vary widely among the others, I modeled its

relationship with online availability by means of negative binomial models (*14*). The negative binomial is a generalization of the Poisson model that allows for an additional source of variance above that due to pure sampling error. A fixed-effects specification of this model refers not to the coefficient estimates but to the "dispersion parameter," forcing the estimated variance of citations to be the same within journals or subfields, but allowing it to take on any value across them. These models were estimated with the Maximum Likelihood method and produced coefficient estimates which, when exponentiated, can be interpreted as the ratio of 1) the number of distinct articles cited following a one year increase in the electronic provision of journals over 2) the number of articles cited without an online increase. One can subtract one from these ratios and multiply by one hundred to get the percentage change of a one-year increase in online availability on the number of distinct items cites. All models contain measures that statistically control for citation year, average number of pages and references in citing articles.

In each subsequent year from 1965 to 2005, more distinct articles were cited from journals and subfields. The pool of published science is growing, and more of it is archived in the CI each year. Online availability, however, has not driven this trend. Fig. 2B illustrates the simultaneous effect of free and online availability on the number of distinct articles cited in journals, and the number of distinct articles and journals cited in subfields. The panels render these effects for journals and subfields with the sample mean of citations. With five additional years of free and commercial online availability, the number of distinct articles cited within journals drops from 600 to 200; articles cited within subfields drops from 25,000 to 15,000; and journals cited within subfields drops from 19 to 16. This suggests that online availability may have reduced the number of distinct articles and journals cited beneath what it would have been had journals not gone online. If a journal provides one additional year of issues online for free,

14% fewer articles are cited. As with citation depth, free availability has a stronger negative association with number of articles cited than commercial provision. The opposite is true, however, for journals cited within subfields. Commercial availability associates with a somewhat greater reduction in journals cited.

Fewer distinct articles and journals are cited soon after they go online. Although this influences the overall concentration of article citations in science, it does not fully determine it. Citations may be spread more evenly over fewer articles to more broadly disperse scientific attention. To assess the degree to which online provision influences the concentration of citations to just a few articles (and journals), a Herfindahl index was computed, where $\sum_{j=1}^{n}(s_j^2)$ represents the percentage of citations *s* to each article *j*, squared and summed across journal or subfield *i* within the 20-year time window examined. A concentration of 1 indicates that every citation to journal *i* in a given year is to a single article; just under this suggests a high proportion of cites pointing to just a few articles; and a concentration approaching zero implies that cites reach out more evenly to a large number of articles. Herfindahl cited articles-in-journal concentrations range from .0000933 to 1 in this sample, with an average of .088 and a wide standard deviation of .195. Where no articles were cited, no concentrations could be computed. Regression models predict citation concentration to articles from the last 20 years with depth of online availability. As in prior models, these were estimated for articles within journals and for articles and journals within subfields, using both commercial and free electronic provision. Citation concentrations are distributed in approximate normality and the models were estimated with OLS.

Figure 2C illustrates the concurrent influence of commercial and free online provision on the concentration of cites to particular articles and journals. The first panel shows that years of

commercial availability significantly increases the concentration of citations to fewer articles within a journal. If an additional 10 years of journal issues were to go online via any commercial source, its citation concentration would rise from .088 to .105, an increase of nearly 20 percent. Free electronic availability has a slight negative effect on the concentration of articles cited within journals, but it has a marginally positive effect on article concentration within subfields, and it appears to significantly drive up the concentration of citations to central journals within subfields. Commercial provision has a much stronger and more consistent positive effect on citation in both articles and journals. For all of the models discussed, similar results were obtained when journals' presence in multiple (e.g., one, two, and three or more) commercial archives was accounted for and modeled simultaneously.

Twenty years is not an unreasonable window within which to examine the effect of online availability on citations, it does not capture the trend of the effect. For example, one can imagine that online provision increases the distinct number of articles cited and decreases the citation concentration for recent articles, but hastens convergence to canonical classics in the more distant past. To explore this possibility, the same analyses were performed with variables calculated using expanding windows ranging from the last year to the last 30 years. To keep samples comparable, all models were estimated on data from 1975 (1945 plus a 30-year window) to 2005 and so the 20-year window coefficients do not correspond perfectly to the effects illustrated earlier. Estimated percentage changes in number of articles and journals cited and the Herfindahl citation concentration within those cites were calculated as associated with a one-year extension of online availability. These estimates and their corresponding 95% confidence intervals are graphed in Figures 2D and E. The top three panels show that increased online provision in the prior year actually lowers the number of distinct articles cited within journals

and articles and journals cited within subfields most in recent years. A one year change in online availability corresponds to a 9% drop in articles cited in the last year, but only a 7% drop in articles cited in the past 20 and 30 years. The pattern is the same for articles and journals within subfields. The three panels of Figure 2D indicate that the citation window's effect on citation concentration is not so consistent. Nevertheless, in the case of article concentrations within subfields, the Herfindahl concentration increase is highest—1.5% per year of online availability—when only calculated for references to last year's articles.

The models presented are limited in a number of ways. For example, journals such as *Science* use Supplemental Online Material for Materials and methods, which frequently include references not indexed by in the CI. It is theoretically possible, though unlikely, that these references are to earlier or more diverse articles. Moreover, by studying only conventional journals, this study fails to capture newer scientific media like science blogs, wikis, and online outlets exploring alternate models of peer review. These new media almost undoubtedly link to extremely recent scientific developments—often through ephemeral weblinks (*15*)—but they may also point to more diverse materials. Nevertheless, this analysis sheds substantial new light on the influence of online availability on the reference patterns in science.

Collectively, the models presented illustrate that as journal archives come online, either through commercial vendors or freely, citation patterns shift. As deeper back-files become available, more recent articles are referenced; as more articles become available, fewer are cited and citations become more concentrate within fewer articles. These changes likely mean that the shift from browsing in print to searching online facilitates avoidance of older and less relevant literature. Moreover, hyperlinking through an online archive puts experts in touch with consensus about what is the most important prior work—what work is broadly discussed and

referenced. With both strategies, experts online bypass many of the marginally-related articles that print researchers skim. If online researchers can more easily find prevailing opinion, they are more likely to follow it, leading to more cites referencing fewer articles. Research on the extreme inequality of internet hyperlinks (*16*), scientific citations (*17*, *18*) and other forms of "preferential attachment" (*19*, *20*) suggest that near-random differences in quality amplify when agents become aware of each other's choices. Agents view other's choices as relevant information—a signal of quality—and factor them into their own reading and citation selections. By enabling scientists to quickly reach and converge with prevailing opinion, electronic journals hasten scientific consensus. But haste may cost more subscription to an online archive.

This research ironically intimates one of the chief values of print library research: poor indexing. Poor indexing—indexing by titles and authors, primarily within core journals—likely had unintended consequences that assisted the integration of science and scholarship. By drawing researchers through unrelated articles, print browsing and perusal may have facilitated broader comparisons and drawn researchers into the past. Modern graduate education parallels this shift in publication—shorter in years, more specialized in scope, culminating less frequently in a true dissertation than an album of articles (*21*).

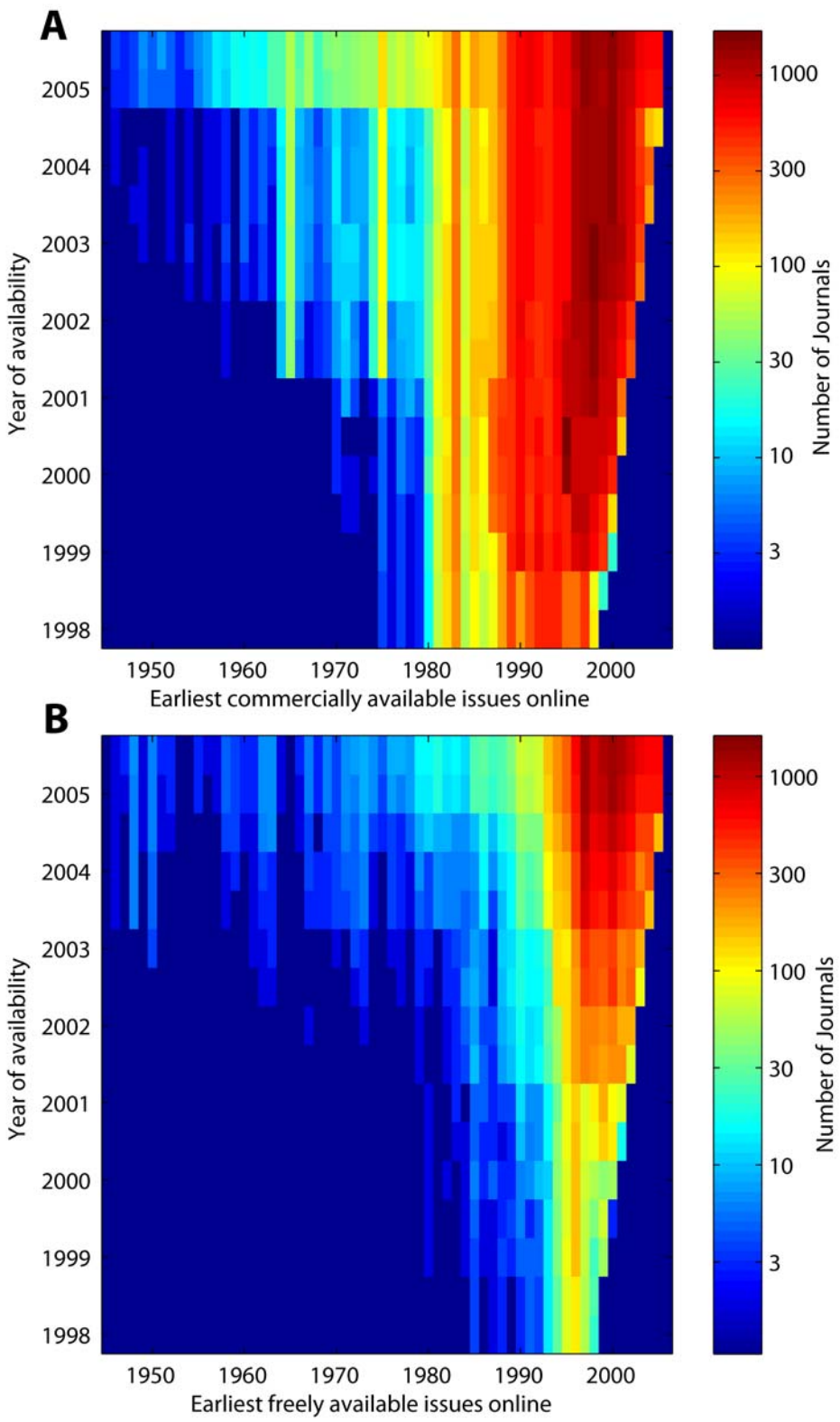
The move to online science appears to represent one more step on the path initiated by the much earlier shift from the contextualized monograph, like Newton's *Principia* (*22*) or Darwin's *Origin of Species* (*23*), to the modern research article. The *Principia* and *Origin*, each produced over the course of more than a decade, were engaged not only in current debates, but wove their propositions into conversation with astronomers, geometers and naturalists from centuries past. As 21st Century scientists and scholars use online searching and hyperlinking to frame and publish their arguments more efficiently, they weave them into a more focused—and more

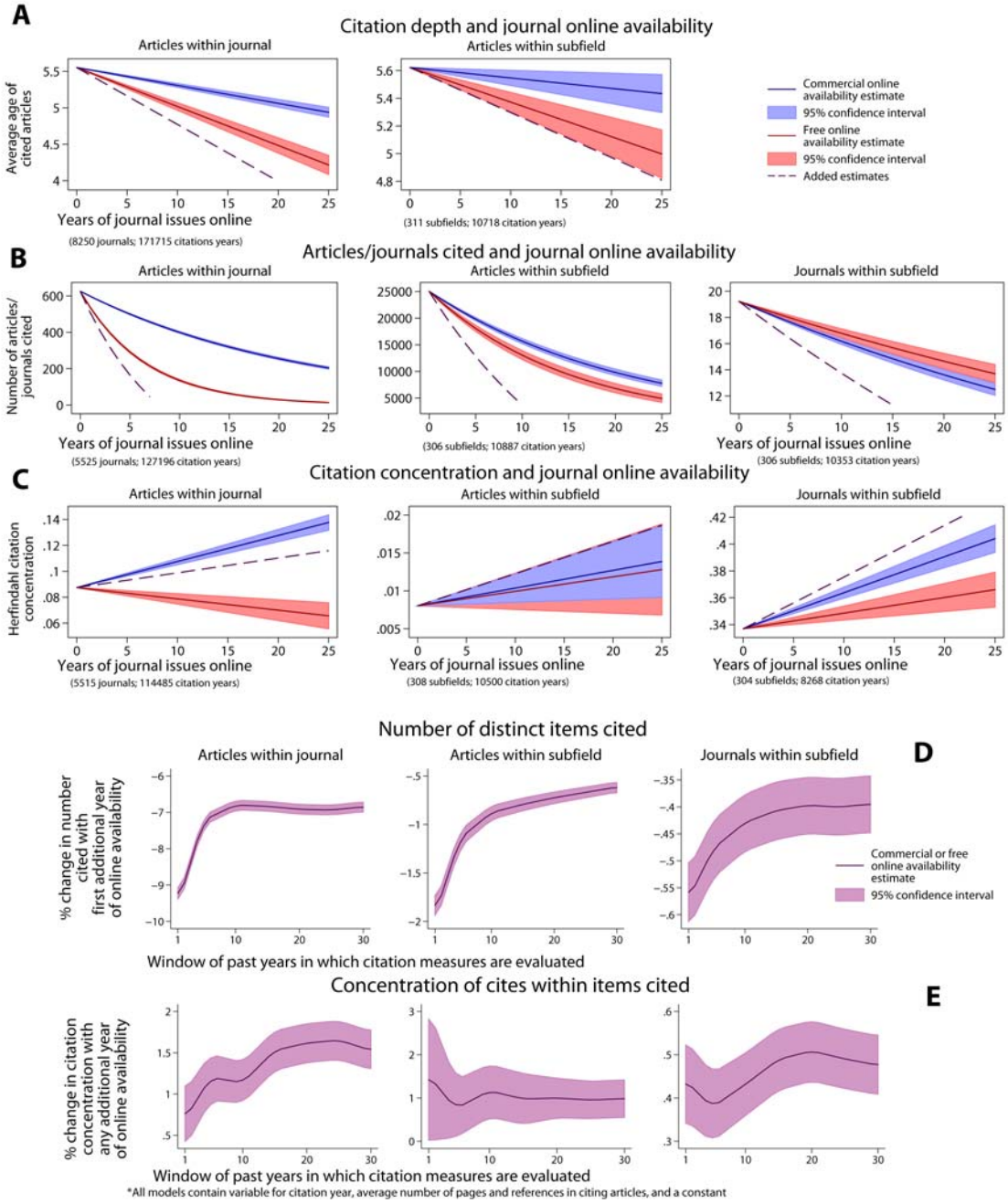
narrow—past and present.

References

1. R. Reddy, I. Wladawsky-Berger, et al. "Digital Libraries: Universal Access to Human Knowledge" (President's Information Technology Advisory Committee, Panel on Digital Libraries, 2001; http://www.nitrd.gov/pubs/pitac/pitac-dl-9feb01.pdf). Note that the report qualifies the vision of universal access, but only by admitting that "more 'quality' digital contents" must be made available and better IT infrastructure must deliver them.
2. S. Harnad, T. Brody. *D-Lib Magazine* **10**, 1082-9873 (2004).
3. M. McLuhan. *Understanding Media*, chapter 1 (1964).
4. S. Black. *Libr. Resour. Tech. Serv.* **49**, 19-26 (2005).
5. S. L. De Groote, J. L. Dorsch. *J. Med. Libr. Assoc.* **91**, 231-240 (2003).
6. C. Tenopir, B. Hitchcock, S. A. Pillow. "Use and Users of Electronic Library Resources: An Overview and Analysis of Recent Research Studies" (Council on Library and Information Resources, 2003).
7. A. Friedlander. "Dimensions and Use of the Scholarly Information Environment: Introduction to a Data Set Assembled by the Digital Library Federation and Outsell, Inc." (Washington, Council on Library and Information Resources, 2002; www.clir.org/pubs/reports/pub110/contents.html).
8. P. Boyce, D. W. King, C. Montgomery, C. Tenopir. *The Serials Librarian*. **46**, 121-141 (2004).
9. C. Tenopir, D. W. King, and A. Bush. *J. Med. Lib. Assoc.* **92**, 233-241 (2004).
10. Shirkey. "Ontology is Overrated: Categories, Links and Tags" (Clay Shirky's 13 Writings About the Internet: Economics & Culture, Media & Community, Open Source. 2005; http://www.shirky.com/writings/ontology_overrated.html).
11. S. Lawrence. Free online availability substantially increases a paper's impact. *Nature*. **411**, 6837: p. 521 (2001).
12. G. Eysenbach. Citation advantage of open access articles. *PLoS Biol.* **4**, e157 (2006).
13. C. Manning and H. Schütz. *Foundations of Natural Language Processing* (MIT Press, Cambridge, 1999), pp. 543.
14. J. Hausman, B. H. Hall, Z. Griliches. *Econometrica* **52**, 909-938 (1984).
15. R. P. Dellavalle, E. J. Hester, L. F. Heilig, A. L. Drake, J. W. Kuntzman, M. Graber, L. M. Schilling. *Science* **302**, 787-788 (2003).
16. A. L. Barabási, R. Albert. *Science*. **286**, 509–512 (1999).
17. R. K. Merton. *Science*. **159**, 56-63 (1962).
18. D. J. Price. *Science* **149**, 510–515 (1965).
19. H. A. Simon. *Biometrika* **42**, 425–440 (1955).
20. M. J. Salganik, P. S. Dodds, D. J. Watts. *Science* **311**, 854-856 (2006).
21. J. Berger. *New York Times* (3 October 2007).
22. I. Newton. *Principia* (Macmillan, New York, ed. 4, 1883). Initially published 1687.
23. C. Darwin. *The Origin of Species* (D. Appleton, New York, 1867). Initially published 1859.

23. I gratefully acknowledge research support from NSF grant 0242971, *Science* Citation Index data from Thompson Scientific, Inc. and Full-text Sources Online data from Information Today, Inc. I also thank Jacob Reimer for helpful discussion and insight.


A
2005
2004
2003
2002
2001
2000
1999
1998
Year of availability
1950
1960
1970
1980
1990
2000
Earliest commercially available issues online
1000
300
100
30
10
3
Number of Journals
B
2005
2004
2003
2002
2001
2000
1999
1998
Year of availability
1950
1960
1970
1980
1990
2000
Earliest freely available issues online
1000
300
100
30
10
3
Number of Journals


A
Citation depth and journal online availability
Articles within journal
Articles within subfield
Average age of cited articles
Years of journal issues online
(8250 journals; 171715 citations years)
(311 subfields; 10718 citation years)
Commercial online availability estimate
95% confidence interval
Free online availability estimate
95% confidence interval
Added estimates
B
Articles/journals cited and journal online availability
Articles within journal
Articles within subfield
Journals within subfield
Number of articles/ journals cited
Years of journal issues online
(5525 journals; 127196 citation years)
(306 subfields; 10887 citation years)
(306 subfields; 10353 citation years)
C
Citation concentration and journal online availability
Articles within journal
Articles within subfield
Journals within subfield
Herfindahl citation concentration
Years of journal issues online
(5515 journals; 114485 citation years)
(308 subfields; 10500 citation years)
(304 subfields; 8268 citation years)
D
Number of distinct items cited
Articles within journal
Articles within subfield
Journals within subfield
% change in number cited with first additional year of online availability
Commercial or free online availability estimate
95% confidence interval
Window of past years in which citation measures are evaluated
E
Concentration of cites within items cited
% change in citation concentration with any additional year of online availability
Window of past years in which citation measures are evaluated
*All models contain variable for citation year, average number of pages and references in citing articles, and a constant

Supporting Online Material

Methods

FSO online availability provides a high resolution picture of journal availability over time. For example, linked issues of FSO show that in the second half of 1998, *Annual Review of Ecology & Systematics* (*ARES*) was not offered online by any major commercial source. The journal first came online in the first half of 1999 through Dialog, with issues going back to January, 1997. ProQuest and H. W. Wilson also began offering the title in the first half of 2000 from 1997. In the second half of 2000, OCLC began to offer *ARES*—the fourth commercial provider, and the journal became available for free through its own website (http://ecolsys.annualreviews.org). A year later, Westlaw and FirstSearch offered it, but Dialog (and Westlaw—which had simply ported Dialog's offerings) stopped offering the most recent year. In the second half of 2002, the journal website pushed online issues back to 1996, and in the first half of 2003, back to November, 1970. In 2002 EBSCO Online—the fifth commercial provider—began offering the journal, and in the second half of 2005, pushed its offerings back through January, 1970. In the first period of 2006, EBSCO and Wilson dropped the most recent three years of the journal.

This *ARES* example shows that the FSO data begins just as library scholars argue the "evolving phase" of online journal availability began—"in the late 1990s," a phase "marked by the availability of both print and electronic journals, with readings from print journals, electronic journals, and journal alternatives" (Boyce, King, Montgomery, and Tenopir 2004). As such, it is appropriate data with which to test propositions concerning online availability on citation patterns over time.

In order to test the effect of online availability on citation depth, the following regression model was estimated, using OLS and specifying fixed effects:

$$CiteDepth_{it} = \alpha_{it} + t\delta + Pages_{it}\gamma_0 + Cites_{it}\gamma_1 + TitleAge_{it}\gamma_2 + IssueDepth_{it}\beta + \nu_i + \varepsilon_{it}$$

where $CiteDepth_{it}$ is the average number of years between publication of an article and year t—the year in which citations are measured—for journal or subfield i ; $IssueDepth_{it}$ is a matrix of variables indicating how many years prior to t that one, two or three commercial archives carried journal i (or any journal from subfield i) and for how many years it was available freely; $Pages_{it}$, $Cites_{it}$ and $TitleAge_{it}$ are the average number of pages, cites and age of weighted title words in citing articles; and $v_i + \varepsilon_{it}$ is the residual with v_i representing the journal or subfield-specific component that differs only between and not within journals or subfields.

To test for the effect of online availability on the distinct number of articles cited in journals and research fields, conditional, fixed-effect negative binomial models were used to account for the overdispersion of $ArticlesCited_{it}$ (*1, 2*). Specifically, we estimated the following model for both journals and subfields:

$$\Pr(ArticlesCited_{i1} = y_{i1},..., ArticlesCited_{in_i} = y_{in_i} \mid \mathbf{X}_i, \sum_{t=1}^{n_i} ArticlesCited_{it} = \sum_{t=1}^{n_i} y_{it})$$
$$= \frac{\Gamma(\sum_{t=1}^{n_i} \lambda_{it})\Gamma(\sum_{t=1}^{n_i} y_{it} + 1)}{\Gamma(\sum_{t=1}^{n_i} \lambda_{it} + \sum_{t=1}^{n_i} y_{it})} \prod_{t=1}^{n_i} \frac{\Gamma(\lambda_{it} + y_{it})}{\Gamma(\lambda_{it})\Gamma(y_{it} + 1)}$$

where $ArticlesCited_{it}$ is the number of articles cited in the twenty years prior to citing year t from journal or subfield i, $\sum_{t=1}^{n_i} y_{it}$ is the sum of these counts for i, Γ is the gamma function, $\lambda_{it} = e^{(\mathbf{x}_{it}\beta + offset_{it})}$ and $\mathbf{x}_{it}\beta = \alpha_{it} + t\delta + Pages_{it}\gamma_0 + Cites_{it}\gamma_1 + IssueDepth_{it}\beta$. Negative binomial models were compared with Poisson models to assess whether the dispersion parameter α was significantly greater than 0—the special case in which the negative binomial is equivalent with the Poisson model. The substantial difference (e.g., $\bar{\chi}^2 = 3.9 \times 10^5$, $p < .001$ for the journal-level model predicted by any online provision)

suggests that a Poisson model fits the data poorly and the negative binomial specification is appropriate. Negative binomial models with varying dispersion were also tested against ones with pooled estimates and found significantly different (e.g., $\bar{\chi}^2 = 3.7 \times 10^5$, $p<.001$ for the journal-level model with any online provision).

In order to capture the degree to which citations are concentrated within particular cited articles, we estimate the following model:

$$\begin{aligned} &ArticleCiteConcentration_{it} = \\ &\alpha_{it} + t\delta + Pages_{it}\gamma_0 + Cites_{it}\gamma_1 + IssueDepth_{it}\beta + \nu_i + \varepsilon_{it} \end{aligned} \quad (3)$$

where $ArticleCiteConcentration_{it}$ is a Herfindahl index capturing the concentration of citations within particular cited articles in the twenty years prior to citing year *t* from journal or subfield *i*.

Tables S1-S4

Table S1 lists descriptive statistics for the variables used in each model of the study. The mean of Citation Depth is 5.6 years for journals and subfields, suggesting that across fields, published work cites articles an average of 5.6 years old. An average of 632 and 25,241 articles are cited from each journal and each subfield each year, with maxima of 2,159 and 35,847, respectively. Not surprisingly, the distribution of citations across journals is more uneven than that across subfields. The mean of article citation in journals and subfields is .088 and .008, both with a much larger standard deviation. Citation concentration to journals within subfields is much higher (.337) and more stable. Citing articles contain are, on average, 12.5 pages in length with 42 references. Citing articles also use overwhelmingly conventional title words—words that have long

appeared in the scientific literature—with an average weighted age of 19 years using a 20 year window. Other independent variables in the analysis underscore the patterns highlighted by Figure 1 in the paper—many more journals are available online through commercial sources than freely, and commercially sources do not all pick up journals at the same time. One will get it, followed gradually by others.

To detail the pattern of online influence across subjects, the 311 fine-grained CI subject codes were consolidated into 14 broad areas: 1) Interdisciplinary Sciences, 2) Biology, 3) Medicine, 4) Environmental & Earth Sciences, 5) Agriculture, 6) Chemistry, 7) Physical Sciences, 8) Engineering, 9) Mathematics, 10) Computer Science & Information Technology, 11) Social Sciences, 12) Business & Management, 13) Law, and 14) Humanities. Tables S2 through S4 illustrate how consistent the effect of years of online availability are on citation depth, distinct articles cited, and article citation concentration across the scientific and scholarly subfields. Journal-level analyses estimates are presented, which are consistent with those at the level of detailed research subfield. All models account for citation year, and are consistent with model presented in the paper which account for age of title words and page and reference numbers from citing articles.

Table S2 shows that the OLS coefficient estimate for any online availability is negative in every subfield and statistically significant ($p<.01$) in all but the computer science & information technology subfield. Commercial and free availability exhibit slightly greater variation, but underlie the consistency of the pattern. R^2 for the models range from .493 for engineering (739 journals) to .722 for business & management (477 journals). ML incidence rate ratio estimates from negative binomial models predicting

the number of distinct article citations in journals with years of online availability are listed in Table S3. These rate ratios all indicate a negative relationship (an incidence ration less than one) between online availability and distinct articles cited. The largest effect appears in the fields of biology, medicine and agriculture whose incidence ratio estimates go down to .884 (agriculture), indicating a nearly 12% drop (e.g., (.88-1.0)*100) in the number of distinct articles cited for each additional year of online availability. In every field, the influence of free electronic availability is stronger than that of commercial access, going down to .815 in the area of medicine.

Table S4 lists estimates for the journal-level models of citation concentration within articles. These effects reveal a wider variance across subfields. In agriculture, there is no effect and in mathematics and engineering, years of free availability appear to exert a significant negative (.002, $p<.05$) influence on citation concentration. In the remaining 11 fields, however, the effect of years online is associated with a rise in the article citation concentration. R^2 fit statistics for the models range from .042 (law) to .130 (social science), suggesting poorer overall fit than in the citation depth models. Collectively these tables show the general consistency of these trends across areas of science and scholarship, and their greatest prevalence in the life sciences.

Tables

Table S1. Variables in Analysis

Variable	Mean	St.dev.	Min	Max
Journal Analysis				
Citation Depth[a]	5.633	3.380	0	20
Articles Cited[b]	632.125	2159.431	0	59891
Article Citation Concentration[c]	0.088	0.195	.00009	1
Citation year[a]	1991.905	10.246	1966	2006
Years online	1.836	5.217	0	61
Years online in one pay source	1.738	4.974	0	61
Years online in two pay sources	.849	2.741	0	36
Years online in three pay sources	.528	2.199	0	36
Years online free	.529	2.503	0	61
Avg. pages in citing articles	12.543	9.405	0	147
Avg. references in citing articles	42.162	19.990	1	236
Avg. title age in citing articles	19.046	1.253	0	20
Subfield Analysis				
Citation Depth[d]	5.622	2.393	0	20
Articles Cited[e]	25241.32	35847.93	0	452800
Journals Cited[f]	22.246	21.542	0	160
Article Citation Concentration[g]	.008	.0536	.00002	1
Journal Citation Concentration[h]	.337	.254	.021	1
Citation year[d]	1986.96	11.488	1966	2005
Years online	3.715	9.656	0	61
Years online in one pay source	3.478	8.960	0	61
Years online in two pay sources	1.928	4.705	0	36
Years online in three pay sources	1.590	4.194	0	36
Years online free	1.678	6.395	0	61

[a]8250 Journals, 171715 Journal-Cite Years; [b]5515 Journals, 119901 Journal-Cite Years
[c]8253 Journals, 195225 Journal-Cite Years; [d]311 Subfields, 10718 Subfield-Cite Years
[e]306 Subfields, 10887 Subfield-Cite Years; [f]298 Subfields; 8567 Subfield-Cite Years
[g]308 Subfields, 10500 Subfield-Cite Years; [h]308 Subfields, 10308 Subfield-Cite Years

Table S2. OLS Coefficient Esimates for Models of Average Citation Depth in Journals: by Broad Subject

Variables	Years online		Years online, pay		Years online, free		R^2
Multidisciplinary Sciences: 82 journals; 2123 citing years							
Model 1	-0.026	(0.007)**					0.537
2			-0.014	(0.008)†			0.535
3					-0.058	(0.011)**	0.540
4			-0.006	(0.008)	-0.056	(0.011)**	0.540
Biology: 1432 journals; 31187 citing years							
	-0.009	(0.002)**					0.526
			-0.001	(0.002)			0.525
					-0.029	(0.004)**	0.526
			0.004	(0.002)†	-0.031	(0.004)**	0.526
Medicine: 2008 journals; 42577 citing years							
	-0.018	(0.002)**					0.518
			-0.015	(0.002)**			0.517
					-0.028	(0.004)**	0.517
			-0.012	(0.002)**	-0.021	(0.004)**	0.518
Environmental & Earth Sciences: 533 journals; 12841 citing years							
	-0.026	(0.003)**					0.572
			-0.024	(0.003)**			0.572
					-0.055	(0.007)**	0.572
			-0.020	(0.004)**	-0.047	(0.007)**	0.573
Agriculture: 189 journals; 4604 citing years							
	-0.024	(0.007)**					0.549
			-0.013	(0.008)†			0.548
					-0.047	(0.012)**	0.549
			-0.005	(0.008)	-0.045	(0.012)**	0.549
Chemistry: 393 journals; 8924 citing years							
	-0.026	(0.004)**					0.453
			-0.022	(0.004)**			0.452
					-0.039	(0.008)**	0.451
			-0.020	(0.004)**	-0.033	(0.008)**	0.453
Physical Sciences: 570 journals; 12841 citing years							
	-0.023	(0.003)**					0.515
			-0.015	(0.003)**			0.514
					-0.044	(0.006)**	0.515
			-0.013	(0.003)**	-0.042	(0.006)**	0.516
Engineering: 739 journals; 17003 citing years							
	-0.014	(0.003)**					0.493
			-0.013	(0.003)**			0.493
					-0.043	(0.008)**	0.493
			-0.011	(0.003)**	-0.038	(0.008)**	0.494
Mathematics: 333 journals; 7397 citing years							

-0.027	(0.004)**				0.598
		-0.025	(0.004)**		0.597
				-0.041	(0.009)** 0.597
		-0.022	(0.005)**	-0.033	(0.009)** 0.598
Computer Science & Information Technology: 416 journals; 8389 citing years					
-0.001	(0.004)				0.610
		0.010	(0.004)*		0.611
				-0.042	(0.007)** 0.612
		0.015	(0.004)**	-0.046	(0.007)** 0.613
Social Sciences: 1314 journals; 31064 citing years					
-0.052	(0.002)**				0.666
		-0.052	(0.002)**		0.666
				-0.075	(0.005)** 0.663
		-0.045	(0.002)**	-0.044	(0.005)** 0.667
Business & Management: 477 journals; 10204 citing years					
-0.041	(0.003)**				0.721
		-0.042	(0.003)**		0.722
				-0.025	(0.007)** 0.715
		-0.043	(0.003)**	0.006	(0.007) 0.722
Law: 155 journals; 4336 citing years					
-0.063	(0.005)**				0.633
		-0.064	(0.005)**		0.633
				-0.071	(0.017)** 0.620
		-0.062	(0.005)**	-0.043	(0.016)** 0.633
Humanities: 586 journals; 14534 citing years					
-0.119	(0.004)**				0.678
		-0.122	(0.004)**		0.677
				-0.127	(0.008)** 0.664
		-0.114	(0.004)**	-0.086	(0.008)** 0.680

*Coefficient for citation year from .184 to .34; Constant from -671.544 to -360.875; all p<.001.
Standard errors in parentheses
† significant at 10%; * significant at 5%; ** significant at 1%

Table S3. ML Estimated Incidence Ratios for Models of Distinct Articles Cited in Journals: by Coarse Subject

Variables	Years Online	Years Online, Pay	Years online, free
Multidisciplinary Sciences: 54 journals; 1516 citing years			
Model 1	.920 (0.006)**		
2		.913 (0.007)**	
3			.890 (0.011)**
4		.942 (0.006)**	.918 (0.011)**
Biology: 908 journals; 21056 citing years			
	.886 (0.002)**		
		.881 (0.002)**	
			.836 (0.003)**
		.935 (0.002)**	.873 (0.004)**
Medicine: 1386 journals; 31001 citing years			
	.893 (0.002)**		
		.845 (0.002)**	
			.815 (0.003)**
		.947 (0.002)**	.846 (0.003)**
Environmental & Earth Sciences: 304 journals; 7829 citing years			
	.907 (0.003)**		
		.905 (0.003)**	
			.851 (0.006)**
		.943 (0.003)**	.879 (0.006)**
Agriculture: 114 journals; 2927 citing years			
	.884 (0.006)**		
		.880 (0.006)**	
			.831 (0.009)**
		.946 (0.006)**	.862 (0.010)**
Chemistry: 215 journals; 4967 citing years			
	.920 (0.005)**		
		.910 (0.005)**	
			.843 (0.008)**
		.968 (0.004)**	.863 (0.008)**
Physical Sciences: 318 journals; 7349 citing years			
	.911 (0.003)**		
		.919 (0.004)**	
			.855 (0.006)**
		.954 (0.003)**	.874 (0.006)**
Engineering: 439 journals; 10903 citing years			
	.929 (0.002)**		
		.935 (0.003)**	
			.852 (0.005)**
		.965 (0.002)**	.868 (0.005)**
Mathematics: 229 journals; 5337 citing years			

	.908	(0.004)**			
			.923	(0.004)**	
					.836 (0.007)**
			.962	(0.003)**	.850 (0.007)**
Computer Science & Information Technology: 272 journals; 5620 citing years					
	.937	(0.003)**			
			.938	(0.003)**	
					.887 (0.005)**
			.965	(0.003)**	.902 (0.005)**
Social Sciences: 1051 journals; 26412 citing years					
	.924	(0.001)**			
			.926	(0.001)**	
					.826 (0.003)**
			.952	(0.001)**	.847 (0.003)**
Business & Management: 386 journals; 8818 citing years					
	.956	(0.002)**			
			.957	(0.002)**	
					.856 (0.005)**
			.972	(0.001)**	.868 (0.005)**
Law: 141 journals; 4112 citing years					
	.951	(0.002)**			
			.950	(0.002)**	
					.824 (0.013)**
			.955	(0.002)**	.839 (0.012)**
Humanities: 491 journals; 13308 citing years					
	.902	(0.002)**			
			.899	(0.002)**	
					.833 (0.006)**
			.919	(0.002)**	.864 (0.005)**

Citeyear coeefficient estimates ranged from .023 to .052; Constant estimates ranged from -102.7 to -44.424, all p<.001.

Table S4. OLS Coefficient Estimates for Models of Citation Concentration in Journals: by Broad Subject

Variables	Years Online	Years Online, Pay	Years online, free	R^2
Multidisciplinary Sciences: 54 journals; 1393 citing years				
Model 1	0.003 (0.001)**			0.070
2		0.003 (0.001)**		0.072
3			0.001 (0.001)	0.061
4		0.003 (0.001)**	0.000 (0.001)	0.072
Biology: 908 journals; 19772 citing years				
	0.001 (0.000)**			0.075
		0.001 (0.000)**		0.075
			0.001 (0.000)*	0.075
		0.001 (0.000)**	0.000 (0.000)	0.075
Medicine: 1385 journals; 29001 citing years				
	0.001 (0.000)**			0.077
		0.001 (0.000)**		0.077
			0.001 (0.000)*	0.076
		0.001 (0.000)**	0.000 (0.000)	0.077
Environmental & Earth Sciences: 304 journals; 7206 citing years				
	0.002 (0.000)**			0.093
		0.002 (0.000)**		0.092
			0.001 (0.001)*	0.090
		0.002 (0.001)**	0.000 (0.001)	0.092
Agriculture: 114 journals; 2751 citing years				
	0.001 (0.001)			0.053
		0.001 (0.001)		0.053
			0.000 (0.001)	0.053
		0.001 (0.001)	-0.001 (0.001)	0.053
Chemistry: 215 journals; 4579 citing years				
	0.001 (0.001)*			0.058
		0.000 (0.001)		0.057
			0.001 (0.001)	0.058
		0.000 (0.001)	0.001 (0.001)	0.058
Physical Sciences: 318 journals; 6741 citing years				
	0.001 (0.000)**			0.050
		0.001 (0.001)*		0.049
			0.000 (0.001)	0.049
		0.001 (0.001)*	0.000 (0.001)	0.049
Engineering: 439 journals; 9551 citing years				
	0.000 (0.001)			0.084
		0.000 (0.001)		0.084
			-0.002 (0.001)*	0.084
		0.001 (0.001)	-0.002 (0.001)*	0.084
Mathematics: 229 journals; 4971 citing years				

0.000	(0.000)				0.083
		0.000	(0.001)		0.083
				-0.002	(0.001)* 0.083
		0.000	(0.001)	-0.002	(0.001)* 0.084
Computer Science & Information Technology: 272 journals; 5059 citing years					
0.003	(0.001)**				0.056
		0.003	(0.001)**		0.056
				-0.002	(0.001)† 0.053
		0.004	(0.001)**	-0.003	(0.001)** 0.058
Social Sciences: 1051 journals; 24600 citing years					
0.003	(0.000)**				0.130
		0.003	(0.000)**		0.130
				0.001	(0.000)* 0.120
		0.004	(0.000)**	-0.001	(0.000)** 0.130
Business & Management: 386 journals; 8040 citing years					
0.002	(0.000)**				0.093
		0.002	(0.000)**		0.093
				-0.001	(0.001) 0.087
		0.002	(0.000)**	-0.003	(0.001)** 0.094
Law: 141 journals; 3861 citing years					
0.003	(0.000)**				0.059
		0.003	(0.000)**		0.060
				-0.001	(0.001) 0.042
		0.003	(0.000)**	-0.002	(0.001)† 0.060
Humanities: 490 journals; 12114 citing years					
0.006	(0.000)**				0.137
		0.006	(0.000)**		0.137
				0.002	(0.001)** 0.122
		0.006	(0.000)**	0	(0.001) 0.137

Coefficient estimate for citation year ranged from -.009 to -.003; Constant estimates ranged from 5.08 to 18.061, all $p<.001$

Standard errors in parentheses

† significant at 10%; * significant at 5%; ** significant at 1%

Additional References

S1. J. Hausman, B. H. Hall, Z. Griliches. *Econometrica* **52**, 909-938 (1984).

S2. B. H. Hall, Z. Griliches, J. A. Hausman. *International Economic Review* **27**, 265-284 (1986).