Does Online Availability Increase Citations? Theory and Evidence from a Panel of Economics and Business Journals

Mark J Mccabe, Christopher M Snyder

DOES ONLINE AVAILABILITY INCREASE CITATIONS? THEORY AND EVIDENCE FROM A PANEL OF ECONOMICS AND BUSINESS JOURNALS

Mark J. McCabe and Christopher M. Snyder*

Abstract—Does online availability boost citations? Using a panel of citations to economics and business journals, we show that the enormous effects found in previous studies were an artifact of their failure to control for article quality, disappearing once fixed effects are added as controls. The absence of aggregate effects masks heterogeneity across platforms: JSTOR has a uniquely large effect, boosting citations around 10%. We examine other sources of heterogeneity, including whether JSTOR disproportionately increases cites from developing countries or to "long-tail" articles. Our theoretical analysis informs the econometric specification and allows citation increases to be translated into welfare terms.

I. Introduction

COULD an economist quadruple his or her citation count just by publishing in an online journal instead of one available only in print? Our initial interest in this question was prompted by the huge effects of online availability found in previous empirical studies. For example, Lawrence (2001) studied a sample of computer science conference proceedings that exhibited within-proceedings variation in availability, with some articles made available online and the rest only in print. In the average proceedings, online articles received 336% more cites than print.[1]

It would not be surprising if convenient online access to the full text of an article boosts its citations. Enhanced access expedites search, allowing citing authors to identify additional relevant articles, and lowers the cost of acquiring, reading, and ultimately citing the articles so identified. Yet the extraordinary size of the estimated effects in these previous studies prompts suspicion that they are biased upward. A possible source of this bias is that the effect of online or open access is confounded with article quality, which is unobservable to the econometrician and so is an omitted variable. For example, Lawrence (2001) makes no mention of whether articles were randomly selected for online publication. If instead of random articles, the best ones were published online, the 336% effect on citations could just be picking up the difference in the citation rates of leading articles versus others.

In this paper, we take on the econometric challenge of identifying the effect of online access from unobserved quality. We are not merely interested in providing a clever solution to an econometric puzzle. Understanding the market for academic journals is important to scholars because it is the one market in which they function as both producers and consumers.[2] Citations are the currency in this market, the prevailing indicator of the impact of scholars' research, advancing a scholar's prestige as well as salary. If a small change in the convenience of access causes cites to quadruple, then the typical citation may be of marginal value, perhaps padding a citing article's references rather than providing an essential foundation for subsequent research.

The question of whether easier access to articles boosts citations has real policy implications. Scholars and librarians have continued to debate the relative merits of traditional journal pricing versus open access. Journals traditionally charged relatively low author fees but high reader fees (in the form of library subscriptions); open access inverts this pricing scheme by providing readers with free Internet access to articles, making up for the revenue loss by increasing author fees. A recent theoretical literature (McCabe & Snyder, 2005, 2007, 2014; McCabe, Snyder, & Fagin, 2013; Jeon & Rochet, 2010) uses a two-sided-market model to assess which pricing scheme would dominate in equilibrium and which would generate the most social surplus.[3] The answer to both questions hinges on the elasticities of demand on the author and reader sides. If online access quadruples

Received for publication September 8, 2011. Revision accepted for publication September 9, 2013. Editor: Philippe Aghion.

* McCabe: Boston University, University of Michigan, and SKEMA Business School; Snyder: Dartmouth College and National Bureau of Economic Research.

We are grateful to the editor, Philippe Aghion, and the referees for suggestions that substantially improved this paper. We are also grateful for helpful comments from Ajay Agrawal, Margo Bargheer, Irene Bertschek, William Bowen, Erik Brynjolfsson, Liz Cascio, Yan Chen, Paul Courant, Glenn Ellison, Joshua Gans, Patrick Gaule, Avi Goldfarb, Dietmar Harhoff, Justus Haucap, Thomas Hess, Tobias Kretschmer, Ethan Lewis, Jeffrey Mackie-Mason, Karen Markey, Thierry Penard, Martin Peitz, Tim Simcoe, Michael Smith, Doug Staiger, Scott Stern, Don Waters, Michael Ward, Harriet Zuckerman, and Christine Zulehner; from seminar participants at Clemson, JSTOR, LMU Munich, the Mellon Foundation, Ohio State, University of Erlangen-Nuremberg, University of Goettingen, University of Michigan, University of Toronto, Yale, and ZEW Mannheim; and conference participants at the Workshop on Economic Perspectives on Scholarly Communication in a Digital Age (Ann Arbor), ZEW Workshop on the Economics of Information and Communication Technologies (Mannheim), International Industrial Organization Conference (Savannah), and the Conference on the Economics of Intellectual Property, Software and the Internet (Toulouse). We are grateful to Mark Bard, Jamie Bergeson-Bradshaw, Chris Faulkner, Yilan Hu, Ella Kim, Scot Parsley, Steve Prager, and Kyle Thomason for excellent research assistance; the many journal publishers and third-party platforms for assistance in constructing journal online histories; and especially Roger Schonfeld for his help in several phases of the project, including facilitating contacts with content providers and platforms, assisting in data collection, and offering insights on developments in scholarly communication. M.M. gratefully acknowledges support from the University of Goettingen, where he conducted part of this research as a DAAD visiting professor. This research was generously supported by a grant from the Andrew W. Mellon Foundation.

[1] See also Curti et al. (2001), who found that online journals generated 54% more cites per article than print-only journals in a cross section of medical journals published from 1995 to 2000.

[2] See Bergstrom (2001) and Dewatripont et al. (2006) for evaluations of the market for academic journals.

[3] In a so-called two-sided market, a platform intermediates transactions between the two sides, tailoring the price charged to each side to ensure sufficient participation on both sides. Participation on one side can benefit the other side, as the presence of many receivers in a telecommunication network exerts a positive externality on a caller or the presence of many readers of an academic journal can exert a positive externality on an author. See Armstrong (2006) and Rochet and Tirole (2006) for surveys of the theoretical literature on two-sided markets.

The Review of Economics and Statistics, March 2015, 97(1): 144–165

doi:10.1162/REST_a_00437

citations, author demand is likely to be quite inelastic, enough to support the high author fees necessary for open access to be sustainable in long-run equilibrium and enough for this open-access equilibrium to have desirable efficiency properties.[4] But if the citation benefit is low, author demand may be so elastic that open access is unsustainable in equilibrium or socially inefficient. Previous research on the effect of open access on citations finds the same huge results as the literature on online access cited above[5] and likely suffers from the same biases,[6] so our results will have implications for the broader question of the effect of enhancing access (whether by placing content online or reducing the fee to download the online content) on citations.

The results also have implications for understanding the transformative value of new technology on scholarly communication. Our analysis will point to JSTOR as the most important platform for online access to the economics literature in the citation period studied (1995–2005). Based on anecdotal evidence, many economists believe that JSTOR substantially enhanced research productivity just as EconLit did in an earlier period and Google Scholar more recently.[7] We provide the first systematic evidence of the impact of JSTOR on the economic literature. Facilitating scientific communication may have broader social welfare implications to the extent that better communication enhances research productivity, which in turn enhances overall economic productivity (see Freeman, 1994; Dosi, 1988).[8] In the theoretical section we show that increases in cites to an article due to a new technology can be linked to increases in the welfare of the article's scholarly audience.

To provide more detail on the paper's contributions, as the title indicates, this paper has both a theoretical and an empirical component. In the theoretical section, we construct a model of citing-author behavior that helps inform the econometric specification. We derive several propositions of independent interest concerning comparative statics and welfare.

The major contribution of the paper is empirical. We address the econometric challenge of separating the citation effect of online access from unobserved quality effects by assembling a large panel data set of citations from any of the thousands of journals indexed by Thomson ISI between 1980 and 2005 to articles published from 1956 to 2005 in a sample of the top one hundred economics and business journals. We merge in hand-collected data on the date that each journal volume was made available on the Internet and over which platforms (i.e., the journal's own website, JSTOR, or several other major Internet platforms). The panel nature of the data set allows us to control for unobserved quality using fixed effects. The considerable exogenous variation in the date of online access across journals allows us to account for secular trends in citations to various vintages of content in economics and business. Additional exogenous variation in the date of online access across volumes of the same journal allows us to account for the age profile of a volume's cites in a flexible way. It is vital to control for these secular trends and age profiles; otherwise they are easily confounded with the online indicators, which tend to "turn on" in later years and for certain ages of content (e.g., only after an embargo window). As we discuss in the literature review, this form of misspecification plagues several of the more recent articles that attempt to correct for the bias due to unobserved quality using panel data.

Our initial set of results shows that the same huge effects of online access found in the previous literature can be generated if fixed effects capturing the quality level of journal volumes are omitted. Once appropriate fixed effects are included, however, the aggregate result cannot be distinguished from 0. Thus, much of the estimated effect of online or open access from the previous literature can be attributed to bias due to omitted quality. We then go on to show that the absence of an estimated effect at the aggregate level masks substantial heterogeneity across platforms. While we find no effect for, among others, Elsevier's ScienceDirect platform, we find a positive and significant effect for JSTOR, boosting citations roughly 10% on average.

We investigate other sources of heterogeneity in the online-access effect—for example, whether the effect differs across high- versus low-ranked journals and whether the effect differs for citing authors in different ranked institutions or different countries. The regional analysis allows us to test the claim by some policymakers that facilitating access should benefit citing authors more in developing countries, where library resources may otherwise be limited.

Section 7 extends our investigation of heterogeneity to the article level by examining whether different articles within a journal volume benefit more or less from online access. Are the effects of online access concentrated among the most cited articles—the "superstars"—or the least cited ones—the "long tail"? Recent studies of online retailing suggest that the latter outcome is predominant: long-tail effects have been found in other markets ranging from

[4] Author fees can be substantial. Currently, the Public Library of Science (PLoS) charges an author fee of $2,900 for *PLoS Biology*, the highest-ranked journal in the ISI biology category.

[5] For example, Harnad and Brody (2004) studied the citation rates of published physics articles, some of which were also self-archived by the author on arXiv (a large online repository offering free downloads of scientific manuscripts). Self-archived articles averaged 298% more cites than the others. Walker (2004) studied an oceanography journal that allowed authors to buy open access for their articles, finding 280% more downloads for open access articles. See Craig et al. (2007) for a survey of research on the citation boost from open access.

[6] The decision by an author to self-archive (studied by Harnad & Brody, 2004) or to pay for open access (studied by Walker, 2004) may be correlated with article quality rather than random. Thus, the large boost in citations these studies attribute to open access may be partly or entirely spurious.

[7] Schonfeld (2003) provides a historical account of the creation and evolution of JSTOR.

[8] A growing literature studies the interplay between scientific publication and innovation. Empirical work by Murray and Stern (2007) finds that patenting ideas first published in scientific articles reduces cites to these articles. Fehder, Murray, and Stern (2014) find that this reduction is concentrated early in the life of a journal, fading as a journal develops a reputation for publishing high-quality scholarship. Their findings suggest that strong intellectual property rights may not impede knowledge flows through established two-sided journal platforms. Theoretical work by Gans, Murray, and Stern (2011) considers the strategic trade-offs involved in disclosing new knowledge through publications, patents, or both.

clothing (Brynjolfsson, Hu, & Simester, 2011) to video sales (Elberse & Oberholzer-Gee, 2008). To date, only Evans (2008) examines these issues in the context of scholarly communication, but as we will discuss, there are reasons to doubt the reliability of his results.

II. Review of Recent Literature

Several recent papers attempt to address the bias due to omitted article quality in estimating the effect on citations of online or open access, but introduce their own specification problems. The closest to our approach are two articles in *Science*: Evans (2008) and Evans and Reimer (2009). These papers use the same basic approach as we do to control for quality by using a panel of citations to individual volumes and including volume fixed effects in their econometric model. Unfortunately their econometric models suffer from a different misspecification problem: although certain age and time variables are included, they do not adequately control for citation trends, which continue to bias the coefficients of interest because online availability also varies systematically across age and time. We demonstrate the point concretely in table 3, where, for example, we reproduce a similar estimate to the 26% for economics and business in Evans and Reimer (2009), but then show that this estimate disappears when the full suite of age and time effects is added.[9]

Contemporaneous research by Depken and Ward (2009) concentrates, as we do, on the effect of JSTOR on citations but unlike us uses the citing article as the unit of analysis. They find that an article written at an institution with access to more JSTOR journals tends to cite these journals more (consistent with our positive JSTOR subscription elasticities) and non-JSTOR journals less (consistent with a rivalry among articles for cites).[10] Our research designs are complementary, providing estimates of the benefits of the JSTOR platform to scholars on opposite sides of the two-sided journals market: theirs for scholars as citing readers and ours for scholars as cited authors. Our paper's additional contributions include considering the full range of formal online platforms, breaking the results down along many different dimensions, and providing additional theoretical results.

Two papers provide convincing identification strategies in detailed case studies of individual platforms. Davis et al. (2008) conducted an experiment in which articles from American Physiological Society journals were randomly selected to be openly accessible immediately on publication, the rest receiving the usual treatment of restricted or fee access for the first year. The randomized design solves the problem of separating the open access effect from unobservable quality. The authors found no differences in citations or in the percentage of articles for the two types of access after one year. Gaule and Maystre (2011) examine the effect of open access on citations to articles in the *Proceedings of the National Academy of Sciences* (*PNAS*) as did earlier studies (Eysenbach, 2006; see also Walker, 2004), but they attempt to control for the endogeneity involved in the author's paying the $1,000 charge for open access by using instruments such as whether the article was published in the last fiscal quarter for the author's affiliated institution (under the presumption that research spending is less elastic then because of "use it or lose it" policies). Instrumenting in this way causes the open access effect to fall by 80% and become statistically insignificant.

The "nonresults" from these two studies are consistent with our finding of no aggregate effect.[11] However, our finding of heterogeneous effects for individual platforms (positive for JSTOR but not for other platforms) calls into question the generalizability of studies of isolated platforms. JSTOR may provide a citation boost because of its desirable properties: it is well known among scholars, includes a large number of journals, and archives all past articles for all listed journals. One may expect little citation boost from the more limited American Physiological Society experimental platform, which may not have been well publicized outside the field, made only a small number of journals available, and offered better access for a scattered sample of articles for just one additional year. *PNAS* may not be the best test case given that most citing scholars have access to the journal through their institutions in any event and that the $1,000 author fee only moves the date of online access up by six months, after which all *PNAS* articles are freely available online.

III. Data

Our analysis is based on a sample of 100 journals in economics and business. We focused on these fields because of

[9] Evans (2008) and Evans and Reimer (2009) have a number of other differences from our paper. They study a broader set of disciplines than economics and business and a larger set of journals within economics and business. This forces them to rely on an electronic database, Fulltext Sources Online, for information on online histories for journals in their sample, whereas we use hand-collected and cross-checked data. Our analysis of the Fulltext data suggested it may have value in understanding broad trends in online access, but that there are drawbacks to its use as a regressor: it omits data for the earliest years of online access (1995–1998), contains inaccuracies in online access dates, and omits important channels (including JSTOR).

[10] In Depken and Ward's (2009) specification, the left-hand-side variable (number of cites to JSTOR-accessible journals) is mechanically related to the right-hand-side variable (number of JSTOR-accessible journals), biasing the JSTOR and rivalry effects upward in absolute value. A way around this bias is provided by Parker, Bauer, and Sullenger (2003) and De Groote, Shultz, and Doranski (2005), who study the citing behavior of authors at a single institution over time (Yale and the University of Illinois medical school, respectively), examining whether adding online access at the author's institution caused the author to increase cites to that journal. While this approach removes the mechanical bias in Depken and Ward (2009), it introduces new problems: different journals may have different secular trends in citations, and these secular trends may be correlated with the access status of the journal. Addressing this problem would require including a set of time effects for journals, which cannot be estimated without data from multiple institutions.

[11] While our aggregate result is consistent with these other studies, the domains of our studies differ: we study the effect of online versus print access, whereas they study the effect of free versus paid access for an online journal.

our interest in the journal market in our own discipline and our knowledge of its institutional details. We also believed that it would be a fertile field in which to find digitization effects. "Softer" fields such as history may rely more on books than journals. The resulting low citation counts may be a noisy measure of quality and access. In some "harder" sciences funded by large grants, the cost of accessing print material may present a relatively small barrier, in which case digitization effects may be small. In the "hardest" fields of math and physics, online access effects may be small for another reason: these fields have a tradition of fairly complete access preprint archives; access to the associated published articles may provide little additional citation boost.[12]

We restricted the sample to 100 journals because of the considerable expense and effort involved for each additional journal. Each journal has many volumes (35 on average in our sample) experiencing different patterns of online access, so we will have many more "experiments" than the 100 journals would imply. The selection procedure was designed to achieve two goals. First, we wanted to focus on the most important journals in the discipline. Second, we wanted to ensure that journals available on JSTOR were represented, based on our a priori belief that JSTOR provides a good source of exogenous variation in the timing of online access and in view of the availability of JSTOR subscription information by journal. The sample in fact includes all of the journals in economics and business that had at least some content posted on JSTOR by 2005 (30 in economics and 18 in business). The remaining journals were selected by first ranking them by the standardized ISI yearly impact factors averaged over the period 1985 to 2004 and then selecting the number of top-ranked journals in each subdiscipline so that the ratio of economics to business journals is the same among JSTOR as among non-JSTOR journals.[13] Overall, 63 journals in economics and 37 in business are in the sample. The sample has an extensive representation of digital channels besides JSTOR. For example, thirty journals published by Elsevier are eventually available online via ScienceDirect.

The data set merges citations data together with historical information on online availability. The citations data were acquired from Thomson ISI. For each of the 100 journals in our sample, ISI lists every article published since 1956. Each published article is linked to all cites from all of the over 8,000 ISI-indexed journals for each year from 1980 to 2005. The database includes detailed information on journal and article title, publication date, author name, affiliation, and location for both the citing and the cited articles. To this basic citation data we merged hand-collected information on online availability of the full-text article. We first identified the major third-party aggregators, which, in addition to the journal publisher's own website, may have been a channel of online access.[14] The major aggregators we consider are JSTOR, EBSCO, ProQuest, Ingenta, Gale, OCLC, and DigiZeitschriften. Then we sought to determine the date on which each journal issue was made available online, if at all, through each channel. This was a painstaking process because information is only readily available regarding current online availability, while our study requires the first date of online availability for each volume. To obtain this information, we contacted the publishers and aggregators, cross-checking their reports using libraries' electronic journal catalogs and the "wayback machine" (www.archive.org), which provides regularly archived snapshots of large segments of the web.

The resulting data set from these two sources includes observations for nearly 260,000 individual cited articles, indexed by i. The analysis is ultimately performed at a more aggregate level—the volume—comprising all of the articles a journal publishes in a year. Aggregating in this way reduces the computational burden—the average volume contains 73 articles—without changing the results—the volume-level estimates are numerically identical to the article-level ones because none of the right-hand-side variables vary at the article level within a volume. Let v index a volume, $j(v)$ index the journal title associated with the volume, and $p(v)$ index the year of the volume's publication. Our data set has a panel structure because each volume receives cites each year over our sample period, from 1980 to 2005. Let t index the citation year. Note the distinction between the data set's two time indexes: $p(v)$ indexes the year the *cited* article was published (from 1956 to 2005) and t indexes the year the *citing* article was published (from 1980 to 2005). Because each journal has many volumes, our sample of 100 journals yields over 3,500 volume observations; because each volume can have as many as 26 citation-year observations (one for each year 1980–2005), our panel yields over 60,000 volume-citation-year observations, the basic unit of analysis for our study.

Table 1 provides descriptive statistics for the data set. All journals were founded by 1988; the earliest, the *Journal of Institutional and Theoretical Economics*, began in 1844. The number of citations to a volume in a single year (*CIT*) is 35.7 on average, or about a half a cite per article. The standard deviation (59.4) is huge, as is the range, from 0 to a maximum of 771 (cites in 2004 to the 1982 volume of

[12] To judge the representativeness of our findings, we conducted a similar analysis—not reported here for space considerations—of 100 journals in each of history and biology. The results for JSTOR subscription elasticities were similar to those reported in table 4.

[13] There was little conflict between the selection of journals based on rank versus the selection on JSTOR availability because JSTOR tends to include top-ranked journals. Only two JSTOR journals, *Canadian Journal of Economics* and *Journal of Risk and Insurance*, would not have been selected based on rank alone. Neither was ranked very far below the cutoff for inclusion in our sample; the former ranked 80 and the latter 89 among economics journals.

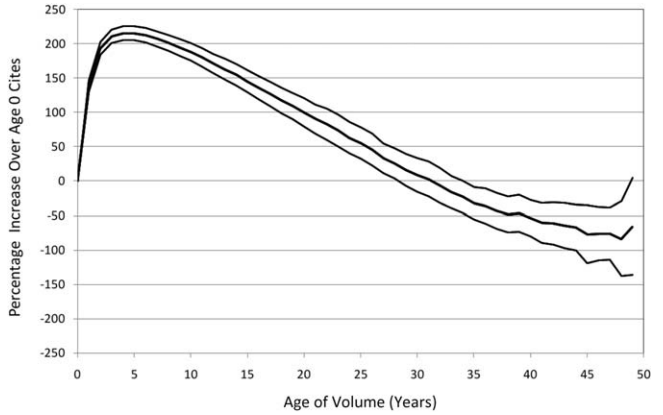
[14] In addition to our own knowledge of the market, we used a number of sources to identify major third-party aggregators, including electronic journal catalogues, for a number of universities and consultations with market experts, one of whom worked for several of the major aggregators.

TABLE 1.—DESCRIPTIVE STATISTICS

	Level of Statistics	Observations	Mean	SD	Minimum	Maximum
Year journal founded	$j(v)$	100	1956.4	28.0	1844	1988
Publication year $p(v)$	v	3,558	1985.7	13.1	1956	2005
Citation year t	vt	60,453	1994.8	7.1	1980	2005
Cites to volume in year *CIT*	vt	60,453	35.7	59.4	0	771
Fraction of volume's articles cited *FCIT*	vt	60,453	0.21	0.21	0	1
Online access indicator *OA*	vt	60,453	0.26	0.44	0	1

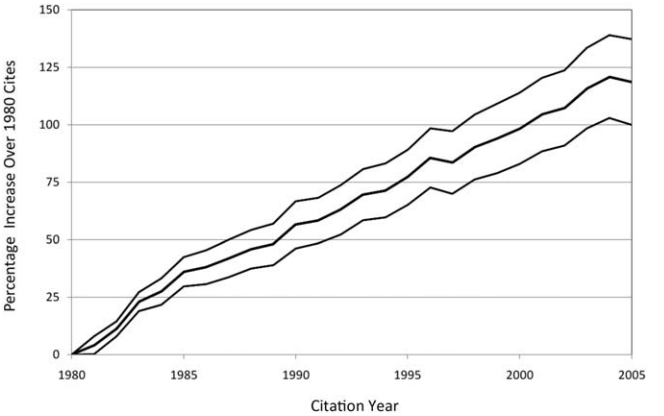
The data set contains journal volumes (indexed by v) observed each year (indexed by t) during the citing period. The journal that publishes volume v is denoted $j(v)$.

FIGURE 1.—CITATION AGE PROFILE



The middle curve plots a set of fixed age effects from Wooldridge's (1999) Poisson quasi-maximum-likelihood procedure. The regression also includes citation-year and journal fixed effects. Outside lines bound the 95% confidence interval based on robust standard errors clustered by journal.

FIGURE 2.—SECULAR TREND IN CITATIONS



The middle curve plots a set of fixed citation-year effects from Wooldridge's (1999) Poisson quasi-maximum-likelihood procedure. The regression also includes a set of age and journal fixed effects. The outside lines bound the 95% confidence interval based on robust standard errors clustered by journal.

Econometrica). The heterogeneity in number of citations that was evident across volumes also is evident within a given volume. Only 21% of the average volume's articles are cited in a given year in the sample. The standard deviation of this measure (0.21) is relatively large, and extreme values are observed (e.g., none of the articles in the 1958 volume of the *Review of Economics and Statistics* were cited in 1991; all of the articles in the 1997 *Quarterly Journal of Economics* were cited in 2003).

Figures 1 and 2 illustrate patterns in citations, which, while interesting in their own right, will be important to account for later in our estimation procedure. Figure 1 plots the profile of citations over the life span of the average journal volume.[15] Citations peak in the fifth year after publication, receiving 216% more than in the year of publication. After that, citations gradually fall each year, falling below the citations received in the year of publication beyond age 30. For the oldest volumes, the age profile asymptotes to about 75% of the cites received in the initial year of publication. The 95% confidence interval shows that the estimates are precise early on in the life cycle but become noisier with age.

Figure 2 plots secular trends in citations. Citations follow a steady upward trend, reaching a level by the end of the sample 120% higher than in the base year of 1980. This increase in citations could be explained by a number of factors, including an increase in the number of citing journals indexed by ISI, an increase in articles per journal, or an increase in the number of references per citing article. In fact, the last factor can account for most of the secular growth in citations as the results, as table 2 shows. The analysis is based on counts of pages and references for a stratified random subsample of articles in our database from each of the 100 journals for the years 1985, 1995, and 2005. The average article in 2005 is 38% longer than in 1985 and has 18% more cites per page, leading to a combined effect of 56% more citations per article. This accounts for almost all of the 61% increase in citations from 1985 to 2005 evidenced in figure 2. The results in figure 2 and table 2 highlight the need to control for secular trends in citations. The results also suggest that any citation boost enjoyed by digitized articles need not have come at the expense of print

[15] Technically, the figure plots the coefficients on a complete set of age indicators from a fixed-effects Poisson regression including fixed effects for journals and citation years. Section V provides a more formal discussion of the estimation procedure, due to Wooldridge (1999).

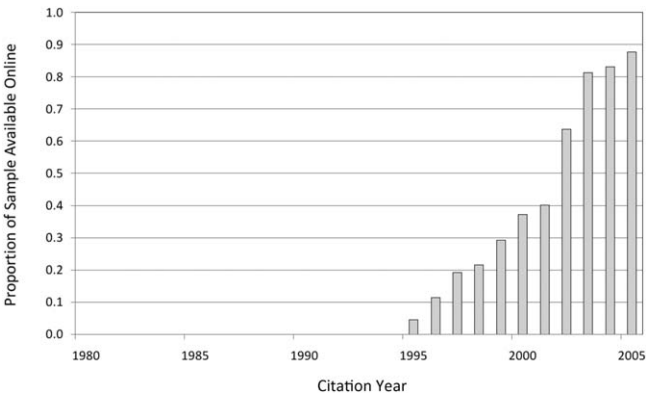
The impossibility of separately identifying age, cohort, and time effects, called the identification problem (Blalock 1966), arises here in that age, volume, and citation-year fixed effects cannot all be separately identified. The age profile is identified in the regressions behind figure 1 by including journal rather than volume fixed effects. In essence, the identifying assumption is that volumes in the same journal have roughly similar citation levels after accounting for time effects. The citation-year profile in figure 2 is identified using a similar strategy. The identification problem will be less of a concern in later regressions because the online access variables of interest are identified after controlling for age, volume, and citation-year fixed effects even though the fixed effects are themselves difficult to identify. See section V, Part C for further discussion.

TABLE 2.—TRENDS IN CITING ARTICLE LENGTH AND REFERENCES

	Pages/ Article	References/ Page	References/ Article
1985–1995 increase	22.1***	11.1**	33.3***
	(3.5)	(4.7)	(4.9)
1995–2005 increase	16.2***	7.4*	22.9***
	(3.8)	(4.5)	(5.3)
Combined 1985–2005 increase	38.3***	18.5***	56.2***
	(4.4)	(4.4)	(5.4)

Analysis on stratified random subsample of five articles from each of the 100 journals in our data set for each year 1985, 1995, and 2005. To omit notes, reviews, and other nonstandard articles, the initial sample was restricted to articles four or more pages long. Dropping remaining nonstandard articles resulted in 1,425 observations. The natural log of the dependent variable in the column heading is regressed on indicators for timing of publication and fixed journal effects. The boxes indicate results from same regression. The combined 1985–2005 increase was estimated from a regression respecifying the indicators for timing of publication. Robust standard errors clustered at the journal level are reported in parentheses. Significantly different from 0 in a two-tailed test at *10%, **5%, ***1%.

FIGURE 3.—TRENDS IN ONLINE AVAILABILITY



The mean (across volumes in sample available to be cited in given year) of the indicator for online access.

articles out of some fixed citation pie but could have been part of a general expansion.

The last row in table 1 provides information on the online access indicator, *OA*. For 26% of the observations, the full volume was available online through some channel for the full year. We will focus on full online access defined in this way throughout the analysis; the regressions will also include indicators for partial online access—only part of a volume's content available online during the year or all of its content available for only part of the year—but we will not focus on those results because partial access is a catch-all category combining observations with various degrees of online access. Figure 3 shows how the full-online-access indicator evolves over time. Full-text articles started to be posted online in 1995. Online access became ubiquitous by the end of our sample, with 88% of volumes available online in 2005. The considerable variation in the online availability of different volumes between 1995 and 2005 will help identify online access effects from secular trends and age effects.

IV. Theoretical Model

In this section we construct a theoretical model of citing-author behavior. We use the model to derive comparative statics results and results linking increases in cites to increases in welfare.

Let I_t be the total number of economics articles available to be cited in year t and $i \in \{1,\ldots,I_t\}$ index a representative article. Similarly, let N_t be the total number of authors who will be citing the economics literature in year t and $n \in \{1,\ldots,N_t\}$ index a representative citing author. A cite from author n to article i results from a two-stage process. In a first stage, he or she becomes aware of the potential value of a set of articles for his or her research through search. In a second stage, for each of the articles of which he is aware, the author decides whether to expend the effort to acquire a full-text copy for use in composing the citing passage.

Online access boosts cites in two ways in the model. First, a citing author may be more likely to find a relevant article during the search stage if the article is available on an online platform.[16] Second, online access lowers the cost of obtaining the full-text article in the acquisition stage, increasing the chance that the quality of the match between the article and a citing author's interests is sufficient to justify the citing author's paying the acquisition cost, which ultimately leads to cites for a greater range of qualities of online articles than print.

Because the search and acquisition effects work in the same direction, the model is able to deliver the straightforward comparative-statics result in proposition 1 that online access weakly increases cites and citing-author welfare under quite general conditions. More subtle is the result that the increase in cites provides a lower bound on the increase in citing-author welfare. Proposition 2 shows this result holds under more restrictive conditions than in proposition 1 but still fairly general conditions. The end of the section pursues a number of extensions and details the empirical implications of the model.

A. Search Stage

With probability $l_{nit} \in [0,1]$, author n learns of the existence of article i and its match quality $q_{nit} \geq 0$ with her research project. Match quality is an amalgam of a vertical dimension (a dimension of quality over which all authors agree that "more is better," such as rigor of the theoretical analysis, precision of the estimates, or novelty of the results) and a horizontal dimension (a dimension over which different citing authors have different opinions, such

[16] While online platforms such as JSTOR undoubtedly facilitate search, given the existence of other powerful search engines—EconLit, which was available from the start of the citation years in our sample, complemented by Google in later years—one might not expect facilitated search to be responsible for a large citation boost. In any event, for completeness, the model allows for online access to affect search.

The advent of digital search dates back well before our sample period. Digital bibliographic data for articles was provided to libraries in the 1970s (Lancaster & Neway, 1982). The development of Mosaic and other popular browers in the mid-1990s allowed individuals to conduct literature searches on their own desktops for free rather than in a library for a fee (Tenopir & Neufang, 1995). It was at this time that academic journals began providing Internet access to the full text of some articles.

as topic). To characterize the distribution of match qualities in the population of citing authors, let $f_{it}(q) \in [0,1]$ be the relative frequency of citing authors in year t who have a match quality q with article i. Let $F_{it}(q) = \sum_{x \le q} f_{it}(x)$ be the associated cumulative distribution function and $\overline{F}_{it}(q) = 1 - F_{it}(q) = \sum_{x>q} f_{it}(x)$ be its complement.

B. Acquisition Stage

In the second stage, citing author n decides whether to acquire full-text copies of the articles in $\{1,\ldots,I_t\}$ of which she is aware, comparing the benefit of acquisition given by the article's match quality q_{nit} to her acquisition cost, a_{nit}. She acquires the article if $q_{nit} > a_{nit}$ and does not if $q_{nit} < a_{nit}$. For concreteness, suppose that when she is indifferent ($q_{nit} = a_{nit}$), she does not acquire the article. Assume that each acquired article results in one cite.

C. Benchmark Analysis

To simplify the benchmark analysis, assume that the probability of learning about the existence of an article l_{nit} is constant across citing authors but may differ across articles, access regimes, and years. In particular, $l_{nit} = l^p_{it}$ if article i is available in print in year t and $l_{nit} = l^o_{it}$ if i is available online, where $l^o_{it} \ge l^p_{it}$. The acquisition cost is treated similarly: $a_{nit} = a^p_{it}$ if article i is available in print in year t and $a_{nit} = a^o_{it}$ if i is available online, where $a^o_{it} \le a^p_{it}$.

Let TC^p_{it} be the total number of cites received by article i in equilibrium in year t if it is available in print and TC^o_{it} if it is available online. We have

$$\begin{aligned} TC^p_{it} &= N_t l^p_{it} \overline{F}_{it}\left(a^p_{it}\right), \\ TC^o_{it} &= N_t l^o_{it} \overline{F}_{it}\left(a^o_{it}\right). \end{aligned} \tag{1}$$

Let TW^p_{it} be the equilibrium total welfare generated by article i over the population of citing authors in year t if it is available in print and TW^o_{it} if it is available online. We have

$$\begin{aligned} TW^p_{it} &= N_t l^p_{it} \sum_{q>a^p_{it}} \left(q - a^p_{it}\right) f_{it}(q), \\ TW^o_{it} &= N_t l^o_{it} \sum_{q>a^o_{it}} \left(q - a^o_{it}\right) f_{it}(q). \end{aligned} \tag{2}$$

Comparative statics results are straightforward. Let $\Delta TC_{it} = TC^o_{it} - TC^p_{it}$ and $\Delta TW_{it} = TW^o_{it} - TW^p_{it}$. The next proposition states that moving from print to online access increases both an article's cites and the welfare it generates for citing authors in equilibrium:

Proposition 1. *$\Delta TC_{it} \ge 0$ and $\Delta TW_{it} \ge 0$.*

The proposition is proved in the appendix. Intuitively, a move from print to digital has two effects on article i, improving the chance that other authors learn of its existence and reducing their cost of accessing it, both leading to an increase in cites and welfare.

The next proposition states that under a fairly general condition, we can use the increase in cites to article i when it moves from print to digital access as a conservative bound on citing authors' welfare gain. The required condition is the log concavity of $f_{it}(q)$. As discussed in Lai and Xie (2006), virtually all discrete distributions familiar to economists (e.g., binomial, Poisson) have this property.[17]

Proposition 2. *Suppose $f_{it}(q)$ is log-concave. Then the percent increase in citing authors' equilibrium welfare when article i moves from print to online availability, $\Delta TW_{it}/TW^p_{it}$, is at least as high as the percent increase in i's* equilibrium cites, *$\Delta TC_{it}/TC^p_{it}$.*

The proposition is proved in the appendix. Intuitively, for log-concave distributions of match quality, as the truncation threshold increases, the mass of the truncated distribution becomes increasingly concentrated just above the truncation point. Print articles have a higher acquisition cost than online articles, so the distribution of match qualities for acquired print articles has a higher truncation point. The distribution of match qualities for acquired print articles is closer to the truncation point, generating relatively little surplus net of the acquisition cost because the acquisition cost is the truncation point.

While proposition 2 provides a useful benchmark, some caveats apply. It rests on the assumption that the probability of identifying an article and the cost of acquiring it are both constant across citing authors. Further, it need not hold if the utility function over the sum of match qualities is concave, as in the extension discussed next.

D. Generalizations

This section generalizes in several dimensions, allowing for a concave function for the citing author's benefit—raising the possibility of rivalry in citations, of which Depken and Ward (2009) found evidence—and allowing search and access cost parameters to differ across authors. We start by examining the effect of access on a representative citing author and then aggregate this behavior up to the market level.

Return to the general formulation in which l_{nit} represents the probability that citing author n's search uncovers article i at time t and a_{nit} represents n's cost of acquiring i once identified. Note that the parameters can differ across authors in this formulation. Moving article i from print to online access will be represented by a weak increase in l_{nit} and weak decrease in a_{nit} for all n (both changes are weak because n's library may not happen to subscribe to the online channel carrying i). Let L_{nit} be the indicator function for author n's learning of the existence of article i; L_{nit} is a Bernoulli random variable equaling 1 with probability l_{nit}, distributed independently across n and across i. Let C_{nit} be

[17] Virtually all continuous distributions familiar to economists (e.g., normal, exponential) are log-concave as well (see Bagnoli & Bergstrom, 2005).

the indicator function for n's undertaking the expense of acquiring i, the main endogenous variable in the model. Author n ends up citing article i in year t if $L_{nit}C_{nit} = 1$ and not if $L_{nit}C_{nit} = 0$. For brevity, we collect the acquisition-cost parameters facing author n in the vector $a_{nt} = (a_{nit})_{i=1}^{I_t}$ and further collect the acquisition-cost vectors across authors in the matrix $a_t = (a_{nt})_{n=1}^{N}$. Define the vectors l_{nt}, q_{nt}, L_{nt}, and C_{nt} and matrices l_t, q_t, L_t, and C_t analogously.

Assume that author n's utility is an increasing, concave function of the sum of match qualities and a decreasing, linear function of total acquisition costs:

$$U_n = B_n(L_{nt} \cdot C_{nt} \cdot q_{nt}) - L_{nt} \cdot C_{nt} \cdot a_{nt}, \tag{3}$$

where $B_n^{'} > 0$ and $B_n^{''} \leq 0$ and where the dots represent inner products of the vectors (i.e., $L_{nt} \cdot C_{nt} = \sum_{i=1}^{I_t} L_{nit}C_{nit}$). Author n's equilibrium decision over which articles to cite is the maximizer of equation (3), denoted $C_{nt}^* = \left(C_{nit}^*\right)_{i=1}^{I_t}$.

Total cites received by one of the articles, i, is given by $TC_{it}^* = \sum_{n=1}^{N_t} L_{nit}C_{nit}^*$. Of central interest are comparative statics effects on the expectation $E\left(TC_{it}^* | a_t, l_t, q_t\right)$. The expectation is conditioned on a_t and l_t because these are the central parameters capturing a move from print to online access; it is conditioned on q_t to hold constant the quality of the available content. The expectation is taken over the distribution of L_t, representing constellations of different articles of which each author becomes aware via search. The comparative statics results are provided in the next proposition. The first result in the proposition indicates that our finding in the benchmark model that a move from print to digital availability boosts an article's own cites holds more generally. The second result provides an example in which a concave benefit function induces rivalry between articles for citations.

Proposition 3. *Suppose article i moves from print to online access, represented by a weak increase in l_{nit} and weak decrease in a_{nit}* for all n:

- *Own citation effect: $E\left(TC_{it}^* | a_t, l_t, q_t\right)$ weakly increases.*
- *Rival citation effect: Suppose further there are exactly two articles, i and j, at time t. Then $E\left(TC_{jt}^* | a_t, l_t, q_t\right)$ weakly decreases.*

The proof, which is based on the monotone comparative statics results of Milgrom and Shannon (1994), is provided in the appendix.

To understand the source of the rivalry result, refer back to the benchmark model in which utility was linear in the sum of match qualities. With that specification, the benefit derived from the marginal cited source is independent of the number of inframarginal sources cited, and thus no rivalry exists among cited sources. With a concave benefit function, the marginal benefit of a cited source declines as more cites are added, raising the possibility that acquiring one article will lead citing authors to ignore others. The rivalry result holds in the special case in which the literature contains only two citable articles. With three or more articles, an improvement in the accessibility of one article can strictly boost cites to another. Acquiring article 1 may reduce author n's marginal benefit of acquiring article 2, knocking 2 off the list of articles n acquires. Knocking 2 off the list may raise the marginal benefit of citing article 3. In this indirect way, improvement in access to 1 can increase cites to rival article 3.[18]

Unfortunately, after generalizing the model to allow for author-specific parameters and concave benefits, it is no longer possible to use the change in citations to bound the change in welfare as done in proposition 2. However, we still regard the bound as a useful starting point to discuss welfare effects.

E. Empirical Implications

We conclude the section with a discussion of the implication of the theoretical model for the specification and interpretation of the empirical model.

The model suggests several ways in which estimates of the citation boost from online availability can be confounded. Higher-quality articles will have better distributions of match quality $f_{it}(q)$ (in the sense of first-order stochastic dominance) and consequently receive more cites on average. There is a danger of mistakenly attributing these quality effects to the effect of online access in a cross section if (as is in fact the case) higher-quality journals are more likely to be put online. We control for this quality effect using our panel data by including fixed effects for journal volumes. The distribution of match qualities also likely varies (in a hump-shaped way; see figure 1) with article age. Age effects may be confounded with online access effects because both trend upward with time. In our empirical specification, included journal-volume fixed effects do not capture age effects, so we will be careful to control for age effects with additional variables (sets of quadratic age profiles).

The model also highlights the importance of controlling for time effects. Cites to an article in the model increase with certain market-wide factors in year t, including the number of citing authors N_t and secular improvements in the distribution of match qualities $f_{it}(q)$ (in the sense of first-order stochastic dominance), due perhaps to a change in professional norms toward including more references in each article. The generalized model raised the prospect of rivalry in citations: a secular increase in the population of citable articles can lead to a reduction in cites to a given

[18] Technically U_n is supermodular in C_{n1t} and $1 - C_{n2t}$ but not also in the additional variable $1 - C_{n3t}$ because the interaction between $1 - C_{n2t}$ and $1 - C_{n3t}$ is negative.

It would be possible to generate examples with only positive rival effects even with a concave benefit function if the citing author could iterate between the two stages of search and acquisition. References in acquired articles in the first iteration could be used to identify useful new articles that are acquired in the next iteration. A reduction in the acquisition cost of one article could have positive spillovers for other articles with this mechanism. This is another reason for taking the rivalry result in proposition 3 as a possibility, not a necessity.

article. All of these effects may contribute to secular trends in citations that need to be controlled for to avoid confounding with online effects, which, as we mentioned, also trend with time. We control for secular trends with a rich set of time fixed effects.

The online access effect is intermediated by online channels and subscribing libraries. The nature of these intermediaries will contribute to the magnitude of the online access effect. Author n will experience a reduction in a_{ni} only if his library subscribes to an online channel carrying article i. Thus, the more libraries that subscribe to online channels and the better the attributes of those channels, the more online access will boost cites. An increase in the effectiveness of an online channel can be captured in the model as a greater increase in l_{nit} and reduction in a_{nit} relative to a less effective channel, leading to a greater citation boost from the effective channel. In theory, online access could end up reducing cites if the publisher took the opportunity to raise prices enough to cause a large drop in library subscriptions.

The model identifies other possible sources of heterogeneity in online access effects. The effect of online access may gradually increase as citing authors grow familiar with the use of different online platforms. On the other hand, if by "online access effect" one is referring to access to the published article through a platform such as JSTOR, this effect may decline over time as cited authors increasingly archive preprint versions of their articles for readers to locate using Google or another search engine. Online access to the published version may provide little citation boost if online access to the preprint version provides a good substitute. Consistent with this possibility, we will see evidence of a rapid rise in self-archiving over the past decade (see table 5).

The online access effect may vary with the quality of the article. Seminal articles may be so valuable that they are acquired even if the procedure is fairly costly, so that online access would have little effect for them compared to more peripheral articles. Formally, for seminal articles, most of the mass of match qualities may lie above both the online and print acquisition costs, so $\overline{F}_{it}(a_{it}^{p})$ may not be too far below $\overline{F}_{it}(a_{it}^{o})$, with both being close to 1. On the other hand, authors may exploit online access to access important articles in other fields, a possible reason for the mass of match qualities in the interval (a_{it}^{o}, a_{it}^{p}) to be substantial, and so a possible reason for the citation boost to be biggest for "superstar" articles.

V. Empirical Methodology

A. Panel Count Data Specification

To account for the count data nature of citations in our panel data setting, we use a fixed-effects Poisson estimator with the following conditional mean,

$$E(CIT_{vt}|\alpha_v, x_{vt}) = \exp(\alpha_v + x_{vt}\beta), \tag{4}$$

where CIT_{vt} denotes citations to journal volume v in year t, α_v is a volume fixed effect, x_{vt} is a vector of regressors, and β is a vector of parameters to be estimated. Although the theoretical analysis in the previous section was conducted at the article level, the econometric analysis will use volume-level observations because, as noted, the estimates are numerically identical but involve less computation. Wooldridge (1999) provides a Poisson quasi-maximum-likelihood (PQML) estimator, which, as long as the conditional mean is specified correctly, produces consistent estimates of β for any positive conditional distribution of CIT_{vt} (Poisson, negative binomial, or other).[19]

Including the volume fixed effects α_v in equation (4) helps remove the bias that plagued previous cross-sectional studies of whether articles available online received more cites than others. If higher-quality articles have greater online availability, the online access variable in these previous studies may just be picking up quality differences between online and print-only articles. To be more precise, "quality" here means the vector of match qualities between the cited article and citing authors at the given moment. The volume fixed effects α_v control for any time-invariant aspect of match quality. Aspects of match quality varying with article age—the typical hump-shaped pattern shown in figure 1—are captured by including a flexibly specified age profile,

$$\gamma_{1k}AGE_{vt} + \gamma_{2k}AGE_{vt}^{2}, \tag{5}$$

where $AGE_{vt} = t - p(v)$ is the age of volume v of journal k in the year of citation and γ_{1k} and γ_{2k} are coefficients that are allowed to vary not just across journals but across five blocks of ten publication years for each journal.[20] It is important to include age controls to avoid, for example, confounding the natural peak in citations at age 5 with online access that might have started in that year.

To control for secular trends, we include a set of interactions between citation and publication years in x_{vt}. These are important controls to include because otherwise the strong secular trends observed in figure 2 might be confounded with online availability, which often occurs later in the sample when secular trends are also highest. The set of citation-publication-year interactions is flexible enough to allow each publication year to have a different secular trend and for each secular trend to have an arbitrary pattern.

The most important regressor in x_{vt} is the variable of interest, the online access indicator OA_{vt}, equaling 1 if volume v was available online in citation year t. We focus on the results for full online access, that is, availability of the entire volume's content for the entire year, but also include controls for partial online access.

B. Heterogeneity in Online Access

Intuition suggests (and the theoretical analysis bears out) that the online access effect may exhibit substantial hetero-

[19] We use Simcoe's (2008) implementation of this estimator in Stata.

[20] Experiments with more flexible age polynomials up to a quartic did not appreciably change the results of interest. The quadratic specification allowed convergence for some of the regressions with fewer observations.

FIGURE 4.—MATRIX OF ESTIMATED ONLINE ACCESS EFFECTS

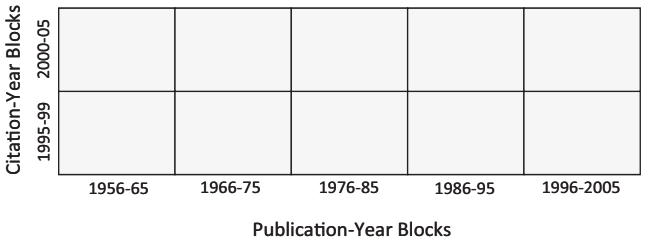
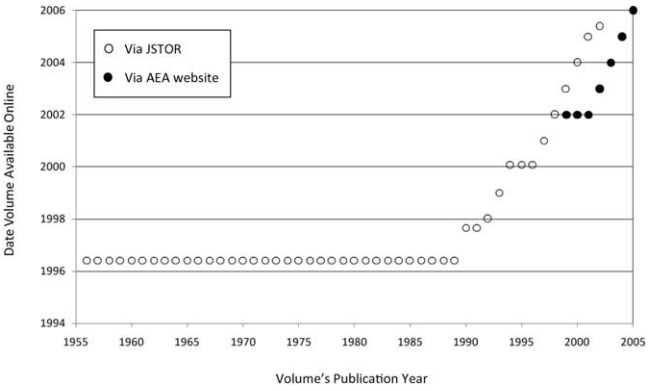


FIGURE 5.—ONLINE AVAILABILITY IN *AMERICAN ECONOMIC REVIEW* EXAMPLE



geneity, varying with time, the characteristics of the channels providing online access, and the inherent quality of the volume. We allow for different effects over time and for different vintages of content by estimating different online access coefficients for each cell in the matrix in figure 4. While our initial regressions consider an aggregate indicator of online access, equaling 1 if the volume was available through any channel, subsequent regressions will allow for different effects across different channels, focusing on the two most important: JSTOR and Elsevier. We will also see if there is any heterogeneity in the individual channel effects across a number of dimensions, including whether online access matters more for popular or unpopular articles, devoting the whole of section VII to this question.

C. *Identification Strategy and Challenges*

Note 15 discussed the identification problem—the impossibility of separately identifying age, cohort, and time effects, arising in many econometric applications. Translated into the present context, the identification problem is the impossibility of separately identifying age, volume, and citation-year fixed effects. Fortunately the problem will not impair our ability to estimate the online access effects of interest. The specified volume, age profile, and citation–publication year interaction variables are not of direct interest themselves but are included as controls to improve the estimation of the online access variables. Estimation of online access effects is not impaired by the identification problem because OA_{vt} varies within these controls.[21]

The variables of interest are not identified if we go as far as to include a different age profile for each volume. It would be impossible to tell if online access was having an effect or if the volume's cites happened to decay more slowly than others for intrinsic reasons. Identification is preserved with more aggregate age profiles; specifically, we specify a profile for each block of ten volumes rather than individual volumes. In essence, our identification assumption is that volumes of a journal that are published around the same time have similar age profiles. If, after netting out own-volume effects and secular trends, we see an increase in citations above this expected citation profile corresponding to when the online access variable turns on, we attribute this effect to online access.

Two challenges must be overcome for the regressor of interest, OA_{vt}, to provide consistent estimates of the online access effect. First, the online access variable must be exogenous in the sense of being uncorrelated with the error term—the difference between the left- and right-hand sides of equation (4). The error mainly consists of the time-varying part of the distribution of match qualities across citing authors that is not picked up by other controls (age profile, interactions between publication and citation year, and so forth). The online access variable will be orthogonal to this error if journals did not look at the pattern of cites to an individual volume and use this information to determine when to place it online. The example of the *American Economic Review* (*AER*), shown in figure 5, suggests that the online access variable is exogenous according to this criterion. The journal was ultimately available online through several channels, including JSTOR and the American Economic Association's website. In 1996, JSTOR placed a whole tranche of volumes online. After that, JSTOR put additional volumes online at the expiration of their "embargo" (the period during which recent content is available only from the publisher, presumably to maintain demand for journal subscriptions). In 2002, the American Economic Association began to make all recent content immediately available online through its own website. This pattern of large tranches together with smaller streams is fairly typical and seems to be based more on technological convenience than on the pattern of an individual volume's citations. It should be emphasized that the only threat to identification is if the journal used time series variation in a volume's cites to decide when to put it online; because of our inclusion of volume fixed effects, there is no threat to identification if information about a volume's mean cites influenced the journal's online posting decision.

The second challenge is that the online access variable must exhibit some independent variation from the other regressors. If volumes were placed online with a fixed lag,

[21] Our primary regressions include a richer set of fixed effects than was used to construct the figures referenced by note 15, including volume rather than more aggregate journal fixed effects, interactions between publication years and citation years, rather than just a set of age and a set of citation effects separately, and five different age profiles for each journal rather than one aggregate age profile.

TABLE 3.—ALTERNATIVE SPECIFICATIONS FOR AGGREGATE RESULTS

Publication Years	Citing Years	(1)	(2)	(3)	(4)	(5)	(6)
1956–1965	1995–1999	5.195***	−0.003	−0.007	0.004	−0.061***	0.010
		(2.177)	(0.071)	(0.055)	(0.033)	(0.013)	(0.026)
	2000–2005	4.589***	0.172	0.123	0.073	−0.007	0.056
		(2.409)	(0.182)	(0.137)	(0.051)	(0.022)	(0.041)
		$n = 8,944$	$n = 8,840$	$n = 8,372$	$n = 8,372$	$n = 8,025$	$n = 8,025$
1966–1975	1995–1999	5.014***	0.033	0.019	−0.010	−0.118***	0.003
		(2.106)	(0.060)	(0.059)	(0.033)	(0.025)	(0.027)
	2000–2005	3.436***	0.067	0.054	0.025	−0.157***	0.032
		(1.526)	(0.094)	(0.090)	(0.024)	(0.021)	(0.029)
		$n = 12,506$	$n = 12,506$	$n = 12,168$	$n = 12,168$	$n = 11,700$	$n = 11,700$
1976–1985	1995–1999	4.085***	0.041	0.051	0.003	−0.155***	0.031**
		(1.773)	(0.083)	(0.086)	(0.017)	(0.016)	(0.015)
	2000–2005	2.505***	0.051	0.052	0.001	−0.190***	0.025
		(1.034)	(0.083)	(0.084)	(0.018)	(0.020)	(0.018)
		$n = 18,441$	$n = 18,441$	$n = 18,394$	$n = 18,394$	$n = 17,618$	$n = 17,618$
1986–1995	1995–1999	0.855***	0.026	−0.027	−0.010	−0.013	−0.004
		(0.324)	(0.044)	(0.021)	(0.020)	(0.017)	(0.017)
	2000–2005	1.709***	0.144***	0.109***	−0.008	−0.039**	0.007
		(0.459)	(0.045)	(0.032)	(0.018)	(0.017)	(0.015)
		$n = 15,062$	$n = 14,907$	$n = 14,789$	$n = 14,789$	$n = 13,815$	$n = 13,815$
1996–2005	1995–1999	0.281	−0.073	−0.143***	0.025	0.019	−0.008
		(0.225)	(0.061)	(0.050)	(0.048)	(0.079)	(0.045)
	2000–2005	1.944***	0.182**	0.004	0.026	0.334***	0.009
		(0.478)	(0.077)	(0.049)	(0.037)	(0.085)	(0.032)
		$n = 5,500$	$n = 5,390$	$n = 5,292$	$n = 5,292$	$n = 4,312$	$n = 4,312$
Fixed effect for source		No	Journal	Volume	Volume	Volume	Volume
Time effects		Full suite	Full suite	Full suite	Full suite	No	Full suite
Quadratic age profile		No	No	No	Yes	No	Yes
Lagged citations		No	No	No	No	Yes	Yes

Results from Wooldridge's (1999) PQML procedure. The dependent variable is cites to a volume in a citing year. Each box reports results of interest from a separate regression for each block of ten publication years. Shown are results for coefficients on the interaction between a full online access variable and two citation-year blocks. Results converted into marginal effects are given by $\exp(\beta) - 1$, where β is the Poisson regression coefficient and $\exp(\beta)$ is the incidence rate ratio. Regressions include online access variables analogous to those reported in the table, but reflecting partial access (access only to part of a volume's content or only for part of the year). The bottom of the table lists other included variables. Full suite of time effects indicates the inclusion of publication-year × citation-year fixed time effects. Robust standard errors clustered at the journal level are reported in parentheses. The number of observations is given at the bottom of each box; some observations may be dropped when moving to a richer specification if cites are constant within a fixed-effect group. Significantly different from 0 in a two-tailed test at *10%, **5%, ***1%.

OA_{vt} would be completely collinear with the volume's age. As figure 5 shows, this is not typically the case. JSTOR began its coverage of the *AER* in 1996 by putting a large tranche (1956–1989) of its back files online. Paradoxically, such tranches help identify the JSTOR effect because simultaneous online availability affects different volumes at different points in their age profiles. For example, JSTOR's initial tranche of over thirty volumes of the *AER* is a shock to the 1956 volume in its fortieth year but the 1957 volume in its thirty-ninth year. The 1956 volume provides information on what the citation age profile should look like up to the thirty-ninth year in the absence of JSTOR. If the 1957 volume deviates from this pattern in 1996, this difference can be attributed to the effect of going online through JSTOR in that year. For this identification strategy to be valid, one must be able to purge secular time effects using data from other journal volumes of around the same vintage having a different pattern of JSTOR availability. Our data satisfy this requirement. Of course, some journals in our sample are never available on JSTOR. For the JSTOR journals, the date that they were introduced to JSTOR varies: the *AER* was introduced to JSTOR in 1996, the *Economic Journal* in 1998, and the *Review of Economic Studies* in 1999.

As figure 5 shows, after the initial tranche of back files, more recent *AER* volumes were added in a stream. However, the stream was not completely regular: JSTOR bunched some recent volumes together as part of its policy to shrink the "embargo" period from five to three years; the American Economic Association inaugurated its own website with three volumes. Any such bunching or, indeed, any small departure from a completely regular annual pattern can be used to identify the effect of online access.

6. Results

A. Alternative Specifications

Table 3 presents the results for our most aggregate measure of the online access effect. All online channels are aggregated; the indicator equals 1 if online access is provided through at least one channel. To demonstrate the importance of the various controls in the preferred specification reported in column 4, the columns leading up to it gradually enrich the included controls. There is some disaggregation allowed even for these aggregate results in that we report the matrix of ten coefficients from figure 4 to allow for heterogeneity in the effect across publication and

citation years.[22] The reported standard errors are robust to heteroskedasticity and clustered at the journal level. Only the results of interest (those for indicators for full online access) are reported; the large number of additional control variables is detailed in the notes for the table. Regression coefficients have been converted into a form interpretable as proportionate increases: a 0 result corresponds to no measured effect from online access, a negative result corresponds to online access causing a reduction in cites, and a positive result corresponds to online access, causing an increase in cites. For example, a result of 0.2 corresponds to cites being 20% higher with online access than without.

An obvious pattern emerges from scanning the table from left to right. Column 1 is run without journal or volume fixed effects to mimic the early literature. Without these controls for quality, we can reproduce the extraordinarily high online access effects found in these studies. For example, the first coefficient of 5.195 has the interpretation that volumes published from 1956 to 1965 received more than a 500%-fold boost in citations from online access in the years 1995 to 1999 compared to having no online access. Similarly huge effects are seen for most of the other entries in the column. The median result for the column is 2.971, representing nearly a 300% boost in citations.

Column 2 adds journal fixed effects. Only two statistically significantly positive results remain, and their magnitudes have been reduced by more than an order of magnitude. Column 3 adds volume fixed effects, an even richer set of quality controls than journal fixed effects, picking up changes in a journal's quality over time.[23] The results are further reduced, and only one statistically significantly positive result remains. The median effect is only 0.05, implying only a 5% citation increase from online access. Column 4 adds a quadratic age profile for each ten-year block of a journal's volumes to the specification in column 3. This further reduces the magnitude of the results toward 0 from both directions. No statistically significant result remains; the median falls to a 1% effect of online access. The fairly tight standard errors suggest precise zero effects from online access at the aggregate level.

Although column 4 reports our preferred specification, we continue with two additional columns of results to provide a formal analysis of the misspecification in two important competing papers surveyed in the introduction: Evans (2008) and Evans and Reimer (2009). Column 5 is our attempt to reproduce Evans and Reimer's (2009) results, which uses the more advanced specification of the two papers.[24] While our underlying data source is different (see note 9), the controls on the right-hand side are the same, including lagged citations and volume fixed effects, which Evans and Reimer included to control for expected citations in the absence of a digitization effect. Importantly, the quadratic age profiles and publication-year × citation-year fixed effects, which we included in all of our specifications to control for secular trends, are absent from column 5. Focus on the last row of results from column 5, because the citation period (2000–2005) is most similar to theirs (1998–2005). We find a greater than 33% boost from online access, similar in magnitude to their finding of about a 26% boost in citations from open access. Thus, in spite of the difference in underlying data, we are able to reproduce their result fairly closely. Column 6 repeats the specification from column 5 but reintroduces the quadratic age profiles and publication-year × citation-year fixed effects. The digitization effect for the 2000–2005 period disappears. This suggests that Evans and Reimer's (2009) results are spurious, generated by omitted-variable bias. Evans and Reimer (2009) do not provide results for earlier periods to which ours can be compared; we find statistically significant and substantial negative digitization effects for these earlier periods in column 5.

We thus see that lagged citations are not an adequate control for time effects. For example, lagged citations do not control for whether citations are rising from the previous year (as they do early in the age profile) or falling (as they do late in the age profile). If these trends are not captured, they will be spuriously picked up in the digitization variable, which also trends over time. This would impart a spuriously positive digitization result for new content and spuriously negative digitization result for old content, exactly what we find in our replication of Evans and Reimer (2009) in column 5: note the negative results for the oldest content in the first four rows and the positive results for the newest content in the last row.

B. *Individual Channels*

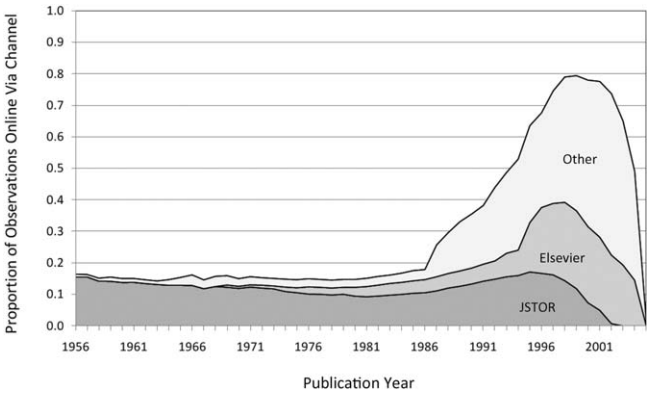
As discussed in sections IV and V, the effect of online access may depend on the nature of the channel providing the access. The number of subscribers to the channel will affect its citation impact, as will its breadth of offerings (a platform with more journals or volumes per journal is more valuable to the researcher), years of operation (users gaining familiarity with the channel over time), and website design.

[22] To allow for considerable flexibility in the coefficients on the included controls, we ran five separate regressions—one for each ten-year block of publication years. Within each regression, the coefficient on online access is allowed to vary across early citation years (1995–1999) and later ones (2000–2005). We use boxes as a device to indicate which results are coming from the same and which from separate regressions.

[23] Engemann and Wall (2009) find considerable movement in their ranking for some journals over shorter periods than our ten publication-year blocks. For example, the *Journal of Industrial Economics* rose sixteen positions between 2002 and 2008, the same number that *Econometric Theory* fell.

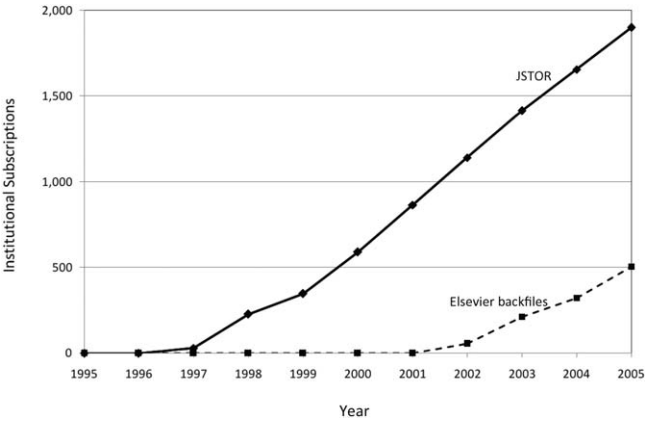
[24] Although not reported in the table, we also estimated a model mimicking the time trend specification in Evans (2008). The regression is similar to that in column 3 except that the quadratic age profile is omitted and time effects are held constant across publication years rather than including the full suite of time effects. The estimated digitization effects are generally higher than in column 3, notably for the lowest entries, with digitization effects as high as 61%, significant at the 1% level. The exercise indicates as much or more positive bias as in Evans and Reimer (2009).

FIGURE 6.—ONLINE ACCESS BY CHANNEL FOR VARIOUS PUBLICATION YEARS



The mean value of indicator of online access via JSTOR or Elsevier for each publication year (including both sole access through these channels and duplicate access through some other channel). Means taken across journals and citation years. "Other" is a residual category equal to the mean of an indicator for online access but not through JSTOR and Elsevier.

FIGURE 7.—SUBSCRIPTION TRENDS



The maximum number of institutions subscribing to an online package provided by the indicated channels containing at least one of the journals in our sample. "Elsevier back files" refers to institutional purchases of its archive of pre-1995 content, which we focus on for Elsevier in our subscription analysis.

Our reported results will restrict attention to the two individual channels covering the most journals (and for which we have subscription data, as will be discussed): JSTOR and Elsevier. Figure 6 shows the proportion of observations in our sample available online through these channels. JSTOR provided online access to a roughly constant fraction of the observations for each publication year, with a fall-off for the most recent publication years as the embargo window was hit. Elsevier gains in importance in more recent publication years, providing online access for fully one-quarter of the sample for publication years 1999 and 2000. About half of the sample journals were carried by JSTOR at some point, and a third were published by Elsevier, so there is a substantial amount of data for each.

Our preferred specification for the individual channel regressions uses institutional subscriptions for JSTOR and Elsevier as a continuous measure of the extent of online access to those channels.[25] For JSTOR, we have data on the number of institutions subscribing to different packages of its online journals. For Elsevier, we have subscriptions for journal back files (volumes of a journal published before 1995) for which an institution could pay a one-time fee for a perpetual access license. We do not have subscription data for Elsevier's more current content, so we restrict the reported Elsevier results to its backfiles.[26] Figure 7 graphs trends in JSTOR and Elsevier back file subscriptions. During 1995–1997, although JSTOR had content posted online, it had few institutional subscribers. After 1997, JSTOR subscriptions increased linearly with time throughout the rest of the sample. Subscriptions to Elsevier back files increased fairly linearly with time after 2001, but the slower rate and later start kept its back files to only about a quarter of JSTOR subscriptions in 2005.

Table 4 reports the results for JSTOR and Elsevier back files subscriptions. For ease of interpretation, the parameters have been converted into elasticities. For example,

TABLE 4.— SUBSCRIPTION ELASTICITIES FOR SELECTED CHANNELS

Publication Years	JSTOR: Sole Channel (1)	JSTOR: With Other Channels (2)	Elsevier Back Files (3)
1956–1965	0.170**	0.138	[a]
	(0.072)	(0.089)	
1966–1975	0.104**	0.186**	0.023
	(0.047)	(0.078)	(0.065)
1976–1985	0.062*	0.068*	−0.049
	(0.036)	(0.036)	(0.054)
1986–1995	0.037	0.062**	−0.012
	(0.027)	(0.025)	(0.047)
1996–2005	0.067***	0.017	[a]
	(0.022)	(0.011)	

Results from Wooldridge's (1999) PQML procedure. The dependent variable is cites to a volume in a citing year. Each box reports the results of interest from a separate regression for each block of ten publication years. Shown are results on the interaction between subscriptions to the channel of interest (JSTOR or Elsevier back files) and an indicator for full online access through the channel; JSTOR results further interacted with indicators for sole access and full access through at least one other channel. Results are converted into elasticities in two steps. First, a marginal effect is computed as $\exp(\beta) - 1$, where β is the Poisson regression coefficient and $\exp(\beta)$ is the incidence rate ratio. Second, the marginal effect is converted into an elasticity by multiplying by the subsample mean number of subscribers. Regressions include the following variables not reported in the table: a complete set of journal-volume fixed effects, a complete set of publication-year × citing-year fixed effects, and a quadratic age profile for each journal. Also included are the following indicators for online access through channels other than JSTOR or Elsevier: partial access only (only access to part of a volume's content or for only part of the year), hybrid access (full access through exactly one channel and partial access through at least one other channel), full access through exactly one channel, and full access through two or more channels. These indicators are aggregated across individual channels and all interacted with two citing-year blocks. Also included is this same set of indicators, but for Elsevier's non-back file (i.e., post-1994) content. Also included are indicators for full online access through JSTOR, interacted with indicators for partial access through some other channel, interacted with JSTOR subscribers. Also included are indicators for partial access through JSTOR interacted with two citing-year blocks. Observations are dropped if among ten or fewer for which any online access indicator is positive. The number of observations is approximately that in column 4 of table 3. Robust standard errors clustered at the journal level are reported in parentheses. Significantly different from 0 in a two-tailed test at *10%, **5%, ***1%.

[a]No observations fit this category.

[25] We also ran an indicator specification consistent with the aggregate regressions in table 3. That is, separate indicators were included for online access through the channel in the early (1995–1999) and late online periods (2000–2005). We also allowed the coefficients to differ depending on whether the channel was the sole source of online access or provided duplicate access. The results were similar to but noisier than the results from the subscription specification so are not reported for space considerations.

[26] We include but do not report indicators for online access to Elsevier's more current content, just as we continue to include indicators for all the other online channels besides JSTOR and Elsevier.

TABLE 5.—TRENDS IN SELF-ARCHIVING OF PREPRINTS

	Subsamples							
	Combined		Published in 1995		Published in 2000		Published in 2005	
	% in Repository	% Solely in Repository	% in Repository	% Solely in Repository	% in Repository	% Solely in Repository	% in Repository	% Solely in Repository
University website	24.6	13.2	10.2	7.4	22.3	12.8	39.7	19.0
SSRN	18.2	7.3	0.0	0.0	20.0	8.9	32.7	12.4
NBER	5.3	0.4	3.2	1.1	6.2	0.3	6.3	0.0
RePEc	1.7	0.9	4.2	2.8	1.0	0.0	0.0	0.0
CEPR	1.8	0.6	2.1	0.7	1.6	0.3	1.6	0.6
Total	35.2	22.5	15.5	12.0	34.8	22.3	53.3	32.1

Analysis on stratified random subsample of five articles from each of the economics journals in our data set for each year 1995, 2000, and 2005. To omit notes, reviews, and other nonstandard articles, the initial sample was restricted to articles four or more pages long. Dropping the remaining nonstandard articles resulted in 903 observations.

the first entry of 0.170 can be interpreted as saying that a doubling of JSTOR subscriptions with online access to the volume would result in a 17.0% increase in citations for that category (sole access through JSTOR to 1956–1965 content). A pair of estimates is provided for JSTOR: one for cases in which JSTOR is the sole channel and another for cases in which JSTOR duplicates the access by some other channel. Because we also include a set of indicators (not reported) for access through other online channels, the JSTOR duplicate estimate should be interpreted as the marginal effect of adding JSTOR access to that provided by these other channels. A single estimate is provided for Elsevier because it is the sole online channel for its back files. The subscription variable already picks up the main trends in citation effects, so we save degrees of freedom by combining citation periods rather than reporting a separate effect for early and late citation periods.

The JSTOR estimates are all positive, and eight of ten are statistically significant. When JSTOR is the sole channel for online access (column 1), a doubling of subscriptions results in an increase in citations of between 3.7% and 17.0%, depending on the publication-year block. The magnitude and statistical significance of the results fall fairly consistently as one moves down column 1 from the oldest to the most recent publication-year blocks. The last row shows a slight reversal in this overall pattern for the most recent content. This reversal should be interpreted with some caution because it is estimated from fewer than forty "treatment" observations for which the sole-JSTOR-availability indicator is turned on—compared to hundreds or thousands of "treatment" observations for the other rows—raising the possibility that a few observations may be unduly influencing the result. For this reason, we will be cautious in interpreting results from the 1996–2005 content in this and the remaining tables in the paper.

Turning to column 2, when JSTOR duplicates access provided by some other online channel, the elasticities are similar to those when JSTOR provides sole online access, ranging between 1.7% and 18.6%. The fact that access to other channels does not seem to impair the marginal contribution of JSTOR access suggests that other channels are not good substitutes for JSTOR. In contrast to the JSTOR results, the Elsevier elasticities are small, sometimes negative, and statistically insignificant. Thus, online access to Elsevier's back file content through its own website (ScienceDirect) appears to provide no citation boost; JSTOR appears to have a uniquely strong effect on citations, and one that is generally strongest for the oldest content.

One explanation for the decline in the JSTOR effect for more recent content is that authors have increasingly resorted to the Internet to distribute their own work, posting pre- or postprint versions of published articles in repositories. The repositories may be accessible enough and the content of the archived article close enough to the published version that there is less citation benefit from having the published version accessible through JSTOR. Our regressions control for online access to the published article online through established channels but not for the author's self-archiving of pre- or postprints (other than what is captured by running separate regressions for citation and publication time blocks).

To investigate the possibility that self-archiving led to the decline in the JSTOR effect, we constructed a stratified random sample of articles from each journal in our data set published in one of three time slices since the penetration of the Internet (1995, 2000, and 2005). We restricted attention to economics journals because of the limited number of well-established repositories in this field: university websites (usually the author's own), the Social Science Research Network (SSRN), the National Bureau of Economic Research (NBER) working paper series, and the Research Papers in Economics (RePEc) and Center for Economic Policy Research (CEPR) archives. For each of the nearly one thousand articles in the sample, we determined whether the article was posted on each of these repositories. Table 5 presents the results. There are two columns for each time slice: the first shows the percentage of articles self-archived on the indicated repository, and the second shows the percentage self-archived solely there, capturing the marginal contribution of the repository to self-archiving. The most important repositories were university websites and the SSRN; around 5% were released as NBER working papers; the remaining repositories were fairly inconsequential. Looking at the overall figures across repositories, only

TABLE 6.—JSTOR SUBSCRIPTION ELASTICITIES BY CITING AUTHOR'S REGION

Publication Years	United States		English-Speaking West		Non-English-Speaking West		Rest of World	
	Sole Access (1)	With Other Channels (2)	Sole Access (3)	With Other Channels (4)	Sole Access (5)	With Other Channels (6)	Sole Access (7)	With Other Channels (8)
1956–1965	0.209**	0.226***	*(0.284)***	*(0.502)****	*(0.121)*	*(−0.085)*	*(0.321)****	*(0.571)****
	(0.092)	(0.092)	*(0.156)*	*(0.186)*	*(0.117)*	*(0.170)*	*(0.121)*	*(0.239)*
1966–1975	0.083*	0.134**	0.046	0.228*	0.027	0.021	*(0.276)****	*(0.509)****
	(0.047)	(0.057)	(0.072)	(0.143)	(0.077)	(0.145)	*(0.107)*	*(0.134)*
1976–1985	0.070	0.098**	0.019	0.106**	0.088	0.010	*(0.160)***	*(0.099)*
	(0.050)	(0.049)	(0.050)	(0.054)	(0.064)	(0.076)	*(0.069)*	*(0.098)*
1986–1995	0.022	0.088***	0.034	0.124**	0.012	0.032	0.088	0.110**
	(0.038)	(0.029)	(0.049)	(0.062)	(0.039)	(0.048)	(0.058)	(0.053)
1996–2005	0.098**	0.000	0.111	0.051*	−0.315*	−0.079**	0.518***	0.047
	(0.049)	(0.016)	(0.103)	(0.032)	(0.147)	(0.045)	(0.128)	(0.065)

The dependent variable is the count of cites from single-authored articles whose author's institution is located in the indicated region. Italics in the boxes indicate regressions that, to obtain convergence, omit the quadratic term in the journal-age profile and retain only the linear term. A country is classified as Western if it is in the United Nations Regional Group "Western Europe and Others" and is classified as English speaking if English is one of its official languages. Specifically, the English-speaking West includes Australia, Canada, Ireland, Israel, New Zealand, and the United Kingdom. For the sake of partitioning the results, this category excludes the United States. The non-English-speaking West includes Austria, Belgium, Denmark, France, Germany, Greece, Italy, Luxembourg, Malta, Monaco, Netherlands, Norway, Portugal, San Marino, Spain, Sweden, Switzerland, and Turkey. Additional specification notes from table 4 apply here.

around 15% of articles published in 1995 were self-archived. The percentage grew over time, so that by 2005, over half were self-archived.[27] The spread of self-archiving is clearly an important change in the publishing environment, which likely reduced the marginal impact of other online channels on citations.

C. Other Sources of Heterogeneity

We next separate the results into finer categories to identify further sources of heterogeneity. We first break the results down by the citing author's country of origin. This breakdown will help us determine whether countries at the center of economics and business research benefit more than the periphery or vice versa. We can test the claim among some policymakers that online access will generate a larger benefit for scholars in developing countries because their libraries have the smallest print collections.

To avoid the ambiguity of defining a location of an article with several authors, we counted cites from single-authored articles only. In this restricted sample, the majority (59%) of cites are from U.S. authors, reflecting the influence of that region in the fields of economics and business. English-speaking Western countries excluding the United States are responsible for 19% of cites, non-English-speaking-Western countries for 16%, and the rest of the world for 6%.

Table 6 presents the regression results for the breakdown by region. Although we include the full suite of online channels used in earlier specifications, in this and subsequent tables we report only the results for JSTOR because of our previous finding that this was the most important individual channel (see table 4). We continue to use the preferred specification involving institutional subscriptions to measure the extent of online access. Each column of boxes is comparable to its analogue in table 4; the only differences are that the left-hand-side variable is the number of citations from single authors in the region rather than in aggregate and that the subscriptions variable on the right-hand side is the number of institutional subscribers located in the given region rather than in aggregate.

Not surprisingly given the proportion of cites coming from the region, the U.S. results in columns 1 and 2 are quite similar to the aggregate ones. JSTOR continues to have a generally positive and statistically significant online access effect of roughly the same size as seen in table 4. The pattern of results is similar for English-speaking Western countries (excluding the United States) in columns 3 and 4.

The results differ for the remaining two regions. In the non-English-speaking West, JSTOR has no statistically significant positive effect. The absence of positive effect does not seem to be due to the relative lack of JSTOR subscriptions in this region: the measure of online access controls for the number of regional subscriptions in this specification; moreover, the number of subscriptions in this region is roughly equal to the number in the English-speaking West and also to the number in the rest of the world. Rather, the lack of a positive JSTOR effect seems instead to be due to greater reliance on national journals not represented in JSTOR by scholars in the non-English-speaking West. Lubrano, Kirman, and Protopopescu (2003) found that the majority of 1991–2000 economics publications in the four largest non-English-speaking European countries appeared in national journals: 66% in Germany, 67% in Spain, 81% in Italy, and 85% in France. Further evidence is provided by Drèze and Estevan (2007), who found that 40 of the 57 journals (85%) appearing on the 2004 National Center for Scientific Research (CNRS) list of top journals ranked by peer opinion in France did not appear on the list of 68 top

[27] For space considerations, table 5 reports only self-archiving of preprints. We also collected information on self-archiving of postprints. An additional 7% of articles were self-archived this way, mostly on SSRN, showing no clear trends over time.

journals ranked by Lubrano et al. (2003) according to an objective citations measure.[28]

The opposite pattern emerges for the rest of the world in the last two columns. The JSTOR subscription elasticities are all positive and generally larger and more statistically significant than those for the United States. The average of the elasticities in columns 7 and 8 is more than double the average in columns 1 and 2. The coefficient in the last row of column 7 implies that a doubling of JSTOR subscriptions would increase citations by 51.8% from the rest of the world. While this increase would not register as a major benefit to the cited author because it is relative to a small base of citations from the rest of the world, it does indicate that the new online technology was a substantial benefit to citing authors from the periphery who likely have less access to print collections, as postulated by some policymakers.

We performed other breakdowns of the result along other possible sources of heterogeneity. One other possible source was the rank of the citing institutions. The idea is that the institutions at the center of citing activity may already have a good infrastructure for their scholars in terms of rich library holdings, research assistance, and administrative support that would allow scholars there to search and acquire relevant articles with either print or online access, whereas scholars in institutions with less support may differentially benefit from online access. To avoid the problem of ranking institutions across different regions of the world, we focused on U.S. institutions only, ranking them by the number of cites coming from single-authored articles from that institution. We again restricted citations to come from single-authored papers to avoid ambiguity in assigning an institutional affiliation to an article with multiple authors. The distribution of citations showed the expected skewness with the top 100 citing institutions responsible for 76% of citations in this sample. The regression results, not reported for space considerations, are fairly similar across the top 100 and other citing institutions. In particular, there is no evidence that more peripheral institutions obtain a disproportionate benefit from JSTOR.

We also broke the results down by the rank of the cited journal, where the same ISI impact factor used in the sample-selection procedure is used here to group the 100 journals into the top and bottom half, again stratified by subfield (economics versus business). The expected skewness in citations again emerges, with the top half of journals in the sample receiving 81% of cites. The regression results, again not reported for space considerations, show some small differences in magnitude and significance between the top half and bottom half of journals, but overall, one category does not appear to obtain a systematically higher citation benefit than the other. A possible explanation for the lack of a systematic difference is that our sample is already restricted to high-impact journals. In any event, we do find that JSTOR provides a citation boost even for the very best journals, ranging as high as an elasticity of 18%.

VII. Long-Tail Effects

The results just mentioned shed light on whether online access effects are journal specific. In this section, we refine the analysis considerably, investigating whether these effects are article specific. The idea that obscure or niche products might disproportionately benefit from Internet search and acquisition was dubbed the *long-tail effect* in Anderson's (2004) famous *Wired* magazine article. In the market for academic journals, the long-tail effect might arise if obscure articles become easier to locate and acquire using the Internet. Seminal articles might experience little effect because they would be well known and important enough to be acquired regardless of the access technology. Such long-tail effects have been recently documented in markets ranging from clothing (Brynjolfsson et al., 2011) to video sales (Elberse & Oberholzer-Gee, 2008). As discussed section IV, the effect could also go the other way, with online access disproportionately boosting citations of the highest-cited articles, sometimes called a "superstar" effect. A superstar effect might arise if online access aids citing authors in identifying and acquiring articles outside their subfields, but only the seminal articles outside of one's subfield are worth citing.

To date, only one paper examines these issues in the context of scholarly communication. Evans (2008) reports that online access reduces the number of cited articles and increases the citation concentration of the articles that are cited, suggesting a superstar effect. By contrast, we will show that the effect of JSTOR access is fairly uniform across the distribution of articles, benefiting superstar and more obscure articles alike. Furthermore, we will find that online access increases the fraction of articles receiving any cites. Our contrasting findings may be due to a number of methodological problems in Evans (2008).[29]

Our approach consists of two complementary estimation strategies. First, we bin the articles into different quintiles based on number of citations received at a certain age and

[28] The only statistically significant results for the non-English-speaking are the negative ones for the most recent content (1999–2005). These results are anomalous because JSTOR availability should at worst be expected to be ignored by citing authors rather than reducing citations. The negative results survived a number of robustness checks, including specifying JSTOR availability as an indicator rather than interacted with subscriptions and including an additional set of publication year × citation year indicators for journals eventually available on JSTOR. As noted above, results in the 1996–2005 row are estimated from many fewer "treatment" observations (for which the content was available solely through JSTOR) than other rows and so should be interpreted with caution.

[29] Evans (2008) specifies journal level rather than volume fixed effects; a single common time trend for all journals is estimated rather than a combination of citation-publication-year interactions and journal-specific quadratic age profiles. As we argued in section VIA, specifications of this sort are unlikely to control for a variety of publication-year and journal-specific time trends, biasing estimates of the parameters that measure the impact of online access. Furthermore, in this 2008 paper, robust standard errors are not estimated in any of the regressions.

TABLE 7.—JSTOR SUBSCRIPTION ELASTICITIES BY QUINTILE

Publication Years	0–20 Quintile		20–40 Quintile		40–60 Quintile		60–80 Quintile		80–100 Quintile	
	Sole Access (1)	With Other Channels (2)	Sole Access (3)	With Other Channels (4)	Sole Access (5)	With Other Channels (6)	Sole Access (7)	With Other Channels (8)	Sole Access (9)	With Other Channels (10)
1956–1965	[a]		[a]		[a]		*(0.323)** *(0.171)*	*(0.322)* *(0.320)*	*(0.136)** *(0.078)*	*(0.105)* *(0.105)*
1966–1975	*0.171* *(0.114)*	*0.222*** *(0.113)*	*0.011* *(0.087)*	*0.290** *(0.173)*	0.309*** (0.089)	0.497** (0.239)	*0.164*** *(0.079)*	*0.200** *(0.110)*	0.101* (0.060)	0.186*** (0.069)
1976–1985	−0.023 (0.080)	−0.034 (0.115)	0.047 (0.046)	0.147** (0.075)	0.049 (0.054)	0.064 (0.107)	0.106** (0.048)	0.107 (0.082)	0.056 (0.040)	0.070 (0.052)
1986–1995	−0.037 (0.052)	0.041 (0.035)	0.036 (0.041)	0.061 (0.040)	0.057 (0.043)	0.026 (0.038)	0.025 (0.035)	0.072** (0.032)	0.028 (0.029)	0.063** (0.032)
1996–2005	0.004 (0.089)	−0.002 (0.033)	0.067 (0.076)	0.017 (0.032)	0.015 (0.081)	0.101*** (0.034)	−0.158*** (0.054)	0.003 (0.022)	0.081** (0.041)	0.015 (0.021)

The quintiles formed by ranking articles within volume by citations in the earliest two citing years available (years used for ranking omitted from regressions). Italics in the boxes indicate regressions that, to obtain convergence, omit the quadratic term in the journal-age profile and retain only the linear term. See table 4 for a list of additional variables included but not reported, notes about number of observations, specification of standard errors, and definition of symbols. The Elsevier back file variables reported in table 4 are also included here but not reported for space considerations.

[a] The regressions had too few observations with positive citations to produce a nonsingular variance-covariance matrix even omitting the linear term in the journal-age profile.

then estimate the online access effect in later years using separate regressions for each quintile. Second, we focus further analysis on the least-cited articles by running regressions involving the proportion of articles in a volume that receive at least one cite. The first approach is described in part A and the second in part B.

A. *Quintile Analysis*

The traditional approach to quintile analysis minimizes a sum of asymmetrically weighted absolute residuals to yield estimates of specific quintiles. While this method has recently been extended to the case of count data (Machado & Silva, 2005), no such estimator has been developed for panel count data. Our alternative consists of applying the Wooldridge (1999) PQML estimator to separate quintiles of articles ranked by the number of citations. In order to avoid bias that can arise if quintiles are based on the contemporaneous value of the dependent variable, which can induce selection-on-the-residuals bias, we use a pre-period of citation years to form the quintile samples but then run the regressions using citation years separated from the pre-period by some gap in time. More specifically, we rank articles by citations in a pre-period window of length t_{vw} years, formed by taking the earliest t_{vw} years of citation data available for that article. The top 20% are placed in the highest quintile group, the next 20% in the next quintile, and so forth. We reaggregate back to the volume level by collecting all the articles within a volume that fall into the same quintile. Notice that the regressions are ultimately run at the volume level, as we have done throughout the analysis, although article-level information was used to form quintiles. The end result is five samples of volume-level data—one subsample for each quintile. We estimate equation (4) separately for each of the five quintile subsamples after discarding observations in the pre-period window along with observations in an additional gap period of t_{vg} citation years. Thus, the regressions are run using only citation years $t_v > t_{vw} + t_{vg}$ for volume v.[30]

The results are reported in table 7. The specification is identical to that in table 4, the only differences being the ones just mentioned: that the aggregated results in table 4 are disaggregated by quintile here and that fewer citation years are used because of the need for a pre-period to form the quintiles. Again, the reported results are converted into subscription elasticities, and just the JSTOR elasticities are reported for space considerations.[31]

The results for the highest-citation (80–100) quintile in columns 9 and 10 are very similar to the corresponding aggregate results in table 4 in both magnitude and significance. Thus, the aggregate results appear to be driven by the most cited articles. The citation boost from a doubling of JSTOR subscriptions when JSTOR is the sole channel for online access ranges from 2.8% to 13.6% and is significant in three of the five publication-year blocks. Similar

[30] Our use of different citation periods for quintile selection and model estimation avoids a bias that would be present with a more naive approach that for each citation year assigns a volume's articles to quintiles based on their observed citation performance for that same citation year. If the regression errors are serially uncorrelated, omitting observations from the pre-period window of t_{vw} citation years will produce consistent estimates. Since we include volume fixed effects in each of the separate quintile regressions, our method will also produce consistent estimates if there is a unit root in the error term for each volume-quintile. The only difficulty that arises for the method is for the intermediate case in which the error term follows an AR(1) process with autocorrelation coefficient $\rho \in (0, 1)$. In this case, omitting the gap period of t_{vg} citation years between quintile selection and estimation will attenuate bias due to selection on the autocorrelated disturbance. While we report results with a two-year window used for quintile selection ($t_{vw} = 2$) and with no additional gap before estimation ($t_{vg} = 0$), as a specification check, we also estimated the regressions using different combinations of selection windows and gaps (t_{vw} ranging from 1 to 3 and t_{vg} ranging from 0 to 4). The results were similar across these alternatives, suggesting that a bias due to an intermediate level of serial correlation in the errors is not a concern.

[31] Online access to Elsevier back files produces no statistically significant citation effects even when broken down at the quintile level, reinforcing the conclusions about Elsevier from table 4.

results are seen for the marginal effect of JSTOR when it duplicates access through other channels. The implication is that the most popular articles receive a citation benefit from JSTOR access.

Looking at the results for lower quintiles in columns 1 to 8, we also see positive and statistically significant effects of JSTOR access for less popular articles. For some blocks of publication and citation years, the results are higher and more significant for lower than the highest quintile, and for others the reverse is true. An examination of the standard errors across each row indicates that the estimates become increasingly noisy as one moves from the highest to the lowest quintile. For the lowest (0–20) quintile, very few results are statistically significant. Still, there is at least some evidence of positive JSTOR effects in all quintiles and nothing to suggest that the proportional effects are greater for the 80–100 quintile. Our interpretation is that positive JSTOR effects can be observed throughout the distribution of articles, from less popular ones in the long tail to the superstars.

TABLE 8.—ELASTICITIES OF PROPORTION OF CITED ARTICLES WITH RESPECT TO SUBSCRIPTIONS TO SELECTED CHANNELS

Publication Years	JSTOR: Sole Channel (1)	JSTOR: With Other Channels (2)	Elsevier Back Files (3)
1956–1965	0.176**	0.200**	[a]
	(0.069)	(0.081)	
1966–1975	0.111***	0.401***	0.071
	(0.039)	(0.056)	(0.085)
1976–1985	0.050*	0.078**	−0.020
	(0.027)	(0.036)	(0.045)
1986–1995	0.030	0.089***	0.065
	(0.029)	(0.032)	(0.042)
1996–2005	−0.034	−0.024	[a]
	(0.033)	(0.026)	

Results from pooled fractional probit estimator developed by Papke and Wooldridge (2008) for panel fractional-response data. The dependent variable is the proportion of articles in a volume cited in a given year. Each box reports the results of interest from a separate regression for each block of ten publication years. Shown are results for an indicator for full online access through the selected channels (JSTOR and Elsevier back files) interacted with subscribers to those channels, separately interacted with indicators for sole access and full access through some other channel. The coefficients are first converted into marginal effects following Papke and Wooldridge's equation (3.10), except we compute the effect at the subsample covariate means rather than computing average partial effects. Marginal effects are then converted into elasticities by scaling by the ratio of subsample means of the independent variable to that of the dependent variable. The subsample mean of subscribers (rather than the interaction of subscribers with online access) is used as a scale factor for the independent variable. See table 4 for the list of the additional variables included but not reported, notes about number of observations, specification of standard errors, and definition of symbols.

B. Fraction of Articles Cited

Given the noise in the estimates for the lowest (0–20) quintile, we take another, complementary approach to studying the effect of online access on the least cited articles, determining if online access affects the proportion of a volume's articles that are cited. Articles that receive no cites in a print world are the true long tail. To quantify the presence of such articles, we go back to the disaggregated article-level data to construct a new variable, $FCIT_{vt}$, measuring the fraction of articles in volume v receiving at least one cite in year t. Descriptive statistics for this variable are provided in table 1.

To deal with a dependent variable having a fractional-response form in a panel data setting with relatively large cross-sectional and small time series dimensions, we employ the pooled fractional probit (PFP) estimator proposed by Papke and Wooldridge (2008), which assumes a conditional mean of the following general form,

$$E(FCIT_{vt}|\alpha_v, x_{vt}) = \Phi\left(\alpha_v + x_{vt}\beta + \overline{x}_{j(v)t}\xi\right), \tag{6}$$

where α_v is a volume fixed effect, here assumed to have a normal distribution conditional on the regressors x_{vt}, Φ is the standard normal cumulative distribution function, $\overline{x}_{j(v)t}$ is the mean value of regressors x_{vt} across volumes for the same journal, and β and ξ are parameter vectors.[32] The results in table 8 have been converted into elasticities to facilitate comparison with our previous results. (See the table notes for details on their construction.)

Like table 4, table 8 reports the results just for the important online channels with subscription data: JSTOR and Elsevier back files. Although involving a different left-hand-side variable than in table 4—fraction cited articles rather than total number of citations—the results are remarkably similar in both size and statistical significance. As column 1 shows, a doubling of JSTOR subscriptions increases the fraction of cited articles by 17.6 percentage points for the earliest content (1956–1965 publication years). This effect gradually becomes smaller with more recent content, but is positive for all but the last block of publication years and statistically significant for the first three blocks. Similar effects are seen in column 2 for JSTOR when it duplicates online access through other channels. Again the effects appear to diminish as the content becomes more recent. Consistent with previous findings, online access to Elsevier back files has no measurable effect on the fraction of cited articles.

Overall, the results from table 8 indicate a significant long-tail effect of JSTOR access. JSTOR access leads to significantly fewer uncited articles. The effect is strongest for the earliest content and gradually disappears for the most recent. These results support the conclusions from the quintile analysis from the previous section that the effects of JSTOR access increase citations throughout the distribution of articles for both popular and obscure ones. By contrast, there is no significant effect of line access to Elsevier back files at any point in the ranking of articles by citations.

[32] The estimator can be implemented in Stata by regressing $FCIT_{vt}$ on a constant, regressors x_{vt}, and regressor means $\overline{x}_{j(v)t}$ using a generalized linear model with a binomial "family" and Bernoulli "link function." Papke and Wooldridge (2008) emphasize the need to cluster the errors at the fixed-effect level, the volume level in our setting. We take a more conservative approach and cluster at the journal level; this also allows us to be consistent with the clustering strategy used previously with the PQML estimator.

VIII. Conclusion

Our empirical analysis of the effect of online availability on cites can be read as a play in two acts. The first act is destructive. By including fixed effects for journal volumes as controls for unobservable quality of the articles in the volume, the estimate of the online access effect was reduced from the extraordinary levels found in the previous literature down to a precisely estimated value of 0. We conclude that the huge estimates found previously are largely spurious, due to these earlier studies' use of cross-sectional data, which prevented them from controlling for unobservable quality. We went on to show that the few studies (e.g., Evans & Reimer, 2009) that attempt to use panel data to get around the bias due to unobservable quality in the earlier literature generally introduce their own specification problem in that they generally lack adequate controls for journal volume age and secular trends in citations. Significant aggregate citation effects disappear when an age profile and a full suite of time effects are included. We conclude that careful specification of the econometric model is as crucial as careful data set construction in identifying the effect of journal access on citations.

The second act is constructive. We show that the zero effect of online access in the aggregate masks substantial heterogeneity across platforms. While some platforms, including Elsevier's ScienceDirect, exhibit no online effects, JSTOR shows significantly positive effects, averaging around a 10% subscription elasticity (meaning that a doubling of JSTOR subscriptions causes a 10% increase in citations). JSTOR has a number of attractive features that may have contributed to its relative importance as a platform: it contains a cross-section of many important journals, it offers access to the entire back file history up to the embargo window, and it was an early entrant in the market. Indeed, JSTOR offered online access to back files five years before ScienceDirect, a long time for users to learn to use the platform and share their experience with colleagues.[33] JSTOR effects tended to be especially large for the earliest content in our sample, that is, articles published between 1956 and 1975. This is consistent with the theoretical model of article search and acquisition in section 4: under a range of conditions, the benefits from online access should be greatest for the content that was heretofore more difficult to access in print. Print access was indeed likely to be more difficult for older content because archival content is often stored in hard-to-access satellite facilities, and EconLit, the major tool for searching the economics literature before Google, did not include information about content published before 1969. We also found that the marginal impact of JSTOR was not diminished if duplicate access was provided by other platforms, such as Ebsco or ProQuest. In most cases, these alternative platforms placed back files online after JSTOR, and often in a piecemeal fashion, likely reducing the relative value of these platforms to citing authors. Overall, while economically meaningful and statistically significant, the JSTOR effect is still modest compared to the huge effects found in the previous literature, which did not control for article quality.[34]

To identify other possible sources of heterogeneity, we disaggregated the results in other dimensions, focusing for the remainder of the paper on JSTOR because this is the channel most likely to give nontrivial results. We found substantial differences in the effect of JSTOR on cites from different regions of the world. Whereas JSTOR had a significant positive effect on citing authors in most other regions, including the United States, it had no effect on the citations from authors in non-English-speaking Europe. One explanation, suggested by the findings of Lubrano et al. (2003), is that authors in this region relied more on national journals rather than the English-language journals available on JSTOR. At the other extreme, JSTOR had around double the effect on citations from the "Rest of the World" (a category including many developing countries) than on U.S. citations. These large effects support the claim by some policymakers that citing authors in developing countries with limited access to print material would benefit more from online access than those in developed countries with extensive libraries.

Other dimensions exhibited less heterogeneity. We hypothesized that articles in lower-tier journals might be more costly and less valuable to access, so an increase in the convenience of access might have a particularly big effect for them. Likewise, we hypothesized that authors from lower-ranked institutions might show a disproportionate increase in citations from online access. Instead we found that the citation boost was fairly uniform across the rank of cited journals and across the rank of the citing author's institution.

We further disaggregated the results by binning the articles into quintiles based on citation rank in a preperiod. We found positive online effects throughout the quintiles. We also found that online access decreases the percentage of articles within a volume that do not receive any cites. Taken together, these results suggest that superstar articles as well as articles residing in the long tail benefit from online access. Thus, the typical power-law relationship between ranked articles and citation counts is shifted up, but its shape is not changed. This result contrasts with studies of long-tail effects in online retail markets, such as books and clothing, where niche products benefit disproportionately

[33] See Harley et al. (2010) for further discussion of the relative merits of JSTOR. Google Scholar, a candidate for the current digital technology with the most impact on scholars, did not appear on the scene until the very end of our sample period, in late 2004.

[34] Our estimates concerning easier access to scholarly literature are lower than recent estimates of the impact of easier access to scientific material. For example, reducing restrictions to genetically engineered mice resulted in 30% more follow-on research (Murray et al., 2009); depositing materials associated with a biomed article in a biological resource center boosted citations to that article by more than 50% (Furman & Stern, 2011).

from use of Internet search capabilities. The difference between the two domains may stem from the different search objectives: whereas retail customers typically search for the single best product match, citing authors search for a bundle of references. Lower-cost access may increase cites to more obscure articles in the author's area of specialization as well as superstar articles outside the author's narrow subdiscipline, simultaneously broadening and deepening the use of the scientific literature.[35]

Tying the results back to the broader policy issues considered in section I, the lack of online access effects at the aggregate level and the modest effects at the channel level resuscitate the view of citations as a valuable currency and useful indicator of an article's contribution to knowledge. At the same time, the modest size of these effects, and the current lack of evidence that free online access performs better, implies that the citation benefits of open-access publishing have been exaggerated by its proponents. Even if publishing in an open access journal were generally associated with a 10% boost in citations, it is not clear that authors in economics and business would be willing to pay several thousand dollars for this benefit, at least in lieu of subsidies. Author demand may not be sufficiently inelastic with respect to submission fees for two-sided-market models of the journal market (e.g., McCabe & Snyder, 2005, 2007, 2014; McCabe et al., 2013; Jeon & Rochet, 2010) to provide a clear-cut case for the equilibrium dominance of open access or for its social efficiency.

The analysis confirms the anecdotal impression that JSTOR was the most important innovation of its time in providing access to the economics and business literature. JSTOR's contribution to social welfare could be substantial. While we do not have empirical evidence directly connecting the measured increase in citations to welfare,[36] our theoretical analysis suggests a connection. Proposition 2 indicates that the 10% citation boost provided by JSTOR on average may (at least in the benchmark model under some conditions) underestimate citing authors' welfare gain from the innovation, which in turn may underestimate the overall social welfare gain to the extent citing authors do not capture the whole social benefit of their articles. The greater boost from JSTOR compared to other channels underscores the value of some of JSTOR's attractive features, including its stability and its coverage of a large number of journals and a complete set of back files for most of these. While they may not revolutionize the scholarly literature, next-generation technologies such as Google Scholar improve on some of these same attractive features and thus promise to continue making measurable contributions to scholarly productivity.

[35] Hervas-Drane (2009) provides a model in which long-tail and superstar effects operate simultaneously in retail markets.

[36] Murray et al. (2009) are able to draw a connection between access to scientific material and scientific progress. See note 34 for further details.

REFERENCES

Anderson, Chris, "The Long Tail," *Wired* 12: 10 (2004).

Armstrong, Mark, "Competition in Two-Sided Markets," *Rand Journal of Economics* 37 (2006), 668–691.

Bagnoli, Mark, and Ted Bergstrom, "Log-Concave Probability and Its Applications," *Economic Theory* 26 (2005), 445–469.

Bergstrom, Theodore, "Free Labor for Costly Journals?" *Journal of Economic Perspectives* 15 (2001), 454–474.

Blalock, Hubert M., "The Identification Problem and Theory Building: The Case of Status Inconsistency," *American Sociological Review* 31 (1966), 52–61.

Brynjolfsson, Erik, Yu (Jeffrey) Hu, and Duncan Simester, "Goodbye Pareto Principle, Hello Long Tail: The Effect of Search Costs on the Concentration of Product Sales," *Management Science* 57 (2011), 1373–1386.

Craig, Iain D., Andrew M. Plume, Marie E. McVeigh, James Pringle, and Mayur Aminet, "Do Open Access Articles Have Greater Citation Impact? A Critical Review of the Literature," *Journal of Informetrics* 1 (2007), 239–248.

Curti, M., V. Pistotti, G. Gabutti, and C. Klersy, "Impact Factor and Electronic Versions of Biomedical Scientific Journals," *Haematologica* 86 (2001), 1015–1020.

Davis, Philip M., Bruce V. Lewenstein, Daniel H. Simon, James G. Booth, and Mathew J. L. Connolly, "Open Access Publishing, Article Downloads, and Citations: Randomised Controlled Trial," *British Medical Journal* 337 (2008), 568–573.

De Groote, Sandra L., Mary Shultz, and Marceline Doranski, "Online Journals' Impact on the Citation Patterns of Medical Faculty," *Journal of the Medical Library Association* 93 (2005), 223–228.

Depken, Craig A., and Michael R. Ward, "Sited, Sighted, and Cited: The Effect of JSTOR in Economic Research," University of Texas at Arlington working paper (2009).

Dewatripont, Mathias, Victor Ginsburgh, Patrick Legros, Alexis Walckiers, Jean-Pierre Devroey, Marianne Dujardin, Françoise Vandooren, Pierre Dubois, Jérôme Foncel, Marc Ivaldi, and Marie-Dominique Heusse, *Study on the Economic and Technical Evolution of the Scientific Publication Markets in Europe* (Brussels: European Commission Directorate General for Research, 2006).

Dosi, Giovanni, "Sources, Procedures, and Microeconomic Effects of Innovation," *Journal of Economic Literature* 26 (1988), 1120–1171.

Drèze, Jacques H., and Fernanda Estevan, "Research and Higher Education in Economics: Can We Deliver the Lisbon Objectives?" *Journal of the European Economic Association* 5 (2007), 271–304.

Elberse, Anita, and Felix Oberholzer-Gee, "Superstars and Underdogs: An Examination of the Long Tail Phenomenon in Video Sales," Harvard Business School working paper 07-015 (2008).

Engemann, Kristie M., and Howard J. Wall, "A Journal Ranking for the Ambitious Economist," *Federal Reserve Bank of St. Louis Review* 91 (2009), 127–139.

Evans, James, "Electronic Publication and the Narrowing of Science and Scholarship," *Science* 321 (2008), 395–399.

Evans, James, and Jacob Reimer, "Open Access and Global Participation in Science," *Science* 323 (2009), 1025.

Eysenbach, Gunther, "Citation Advantage of Open Access Articles," *PLoS Biology* 4 (2006), 692–698.

Fehder, Daniel C., Fiona E. Murray, and Scott Stern, "Intellectual Property Rights and the Evolution of Scientific Journals as Knowledge Platforms," *International Journal of Industrial Organization* 36 (2014), 83–94.

Freeman, Chris, "The Economics of Technical Change," *Cambridge Journal of Economics* 18 (1994), 463–514.

Furman, Jeffrey L., and Scott Stern, "Climbing Atop the Shoulders of Giants: The Impact of Institutions on Cumulative Research," *American Economic Review* 101 (2011), 1933–1963.

Gans, Joshua S., Fiona E Murray, and Scott Stern, "Contracting over the Disclosure of Scientific Knowledge: Intellectual Property and Academic Publication," SSRN working paper abstract 1559871 (2011).

Gaule, Patrick, and Nicholas Maystre, "Getting Cited: Does Open Access Help?" *Research Policy* 40 (2011), 1332–1338.

Harley, Diane, Sophia Krzys Acord, Sarah Earl-Novell, Shannon Lawrence, and C. Judson King, *Assessing the Future Landscape of Scholarly Communication: An Exploration of Faculty Values and*

Needs in Seven Discipline (Berkeley: University of California at Berkeley Center for Studies in Higher Education, 2010), 12, 2013 from http://escholarship.org/uc/cshe_fsc.
Harnad, Steven, and Tim Brody, "Comparing the Impact of Open Access (OA) vs. Non-OA Articles in the Same Journals," *D-Lib Magazine* 10:6 (2004),
Hervas-Drane, Andres, "Word of Mouth and Taste Matching: A Theory of the Long Tail," NET Institute working paper 07-41 (2009).
Jeon, Doh-Shin, and Jean-Charles Rochet, "The Pricing of Academic Journals: A Two-Sided Market Perspective," *American Economic Journal: Microeconomics* 2 (2010), 222–255.
Lai, Chin-Diew, and Min Xie, *Stochastic Ageing and Dependence for Reliability* (New York: Springer, 2006).
Lancaster, F. W., and Julie M. Neway, "The Future of Indexing and Abstracting Services," *Journal of the American Society for Information Science* 33 (1982), 83–89.
Lawrence, Steve, "Free Online Availability Substantially Increases a Paper's Impact," *Nature* 411 (2001), 521.
Lubrano, Michel, Alan Kirman, and Camelia Protopopescu, "Ranking Economics Departments in Europe: A Statistical Approach," *Journal of the European Economic Association* 1 (2003), 1367–1401.
Machado, José A. F., and J. M. C. Santos Silva, "Quantiles for Counts," *Journal of the American Statistical Association* 100 (2005), 1226–1237.
McCabe, Mark J., and Christopher M. Snyder, "Open Access and Academic Journal Quality," *American Economic Review Papers and Proceedings* 95 (2005), 453–458.
——— "Academic Journal Prices in a Digital Age: A Two-Sided Market Model," *B.E. Journal of Economic Analysis and Policy* 7:1 (2007), art. 2.
——— "The Economics of Open Access Journals," Dartmouth College working paper (2014).
McCabe, Mark J., Christopher M. Snyder, and Anna Fagin, "Open Access versus Traditional Journal Pricing: Using a Simple 'Platform Market' Model to Understand Which Will Win (and Which Should)," *Journal of Academic Librarianship* 39 (2013), 11–19.
Milgrom, Paul, and Chris Shannon, "Monotone Comparative Statics," *Econometrica* 62 (1994), 157–180.
Murray, Fiona, Philippe Aghion, Matthias Dewatripont, Julian Kolev, and Scott Stern, "Of Mice and Academics: Examining the Effect of Openness on Innovation," NBER working paper 14819 (2009).
Murray, Fiona E., and Scott Stern, "Do Formal Intellectual Property Rights Hinder the Free Flow of Scientific Knowledge? An Empirical Test of the Anti-Commons Hypothesis," *Journal of Economic Behavior and Organization* 63 (2007), 648–687.
Papke, Leslie E., and Jeffrey M. Wooldridge, "Panel Data Methods for Fractional Response Variables with an Application to Test Pass Rates," *Journal of Econometrics* 145 (2008), 121–133.
Parker, Kimberly, Kathleen Bauer, and Paula Sullenger, "E-Journals and Citation Patterns: Is It All Worth It?" *Serials Librarian* 44 (2003), 209–213.
Rochet, Jean-Charles, and Jean Tirole, "Two-Sided Markets: A Progress Report," *Rand Journal of Economics* 37 (2006), 645–667.
Schonfeld, Roger, *JSTOR: A History* (Princeton, NJ: Princeton University Press, 2003).
Simcoe, Timothy, "XTPQML: Stata Module to Estimate Fixed-Effects Poisson (Quasi-ML) Regression with Robust Standard Errors," Statistical Software Components, Boston College Department of Economics (2008), econpapers.repec.org/RePEc:boc:bocode: s456821.
Tenopir, Carol, and Ralf Neufang, "Electronic Reference Options: Tracking the Changes," *Online* 16 (1995), 67–73.
Walker, Thomas, "Open Access by the Article: An Idea Whose Time Has Come?" *Nature Web Focus* (2004), art. 13.
Wooldridge, Jeffrey M., "Distribution-Free Estimation of Some Nonlinear Panel Data Models," *Journal of Econometrics* 90 (1999) 77–97.

APPENDIX

This appendix provides proofs of the propositions in the text. The first proposition holds for arbitrary distributions; the second requires the additional condition of log-concavity. Bagnoli and Bergstrom (2005) discuss the wide applicability of results related to log concavity in the economics literature. Their familiar results are for continuous random variables and so do not directly apply to our setting, which uses a discrete distribution of match qualities to reflect the discrete nature of the population of authors. We draw on results for log-concave discrete distributions collected in Lai and Xie's (2006) text. The proofs for the case with a continuum of authors and a continuous distribution over match qualities would be similar, based on theorems in Bagnoli and Bergstrom (2005).

To streamline the proofs of the propositions, we provide some technical results as lemmas. The first lemma shows that the complement to the distribution function (also called the survivor function) is nonincreasing:

Lemma 1. *$\overline{F}_{it}(q)$ is nonincreasing in q.*

Proof. Consider two values of match quality, $q'' > q'$. Then $\overline{F}_{it}(q') = \sum_{q\in(q',q'']} f_{it}(q) + \overline{F}_{it}(q'') \geq \overline{F}_{it}(q'')$. *Q.E.D.*

Let $MRL_{it}(q)$ be the mean residual life function:

$$MRL_{it}(q) = \frac{\sum_{x>q}(x-q)f_{it}(x)}{\overline{F}_{it}(q)}. \tag{A1}$$

In our setting, $MRL_{it}(q)$ can be interpreted as the expected increase in match quality moving from an author with match quality q to one with a greater match quality. (See Bagnoli & Bergstrom, 2005, for a further discussion of the interpretation of this function and its wide applicability in economics).

Lemma 2. *$MRL_{it}(q)$ is nonincreasing if $f_{it}(q)$ is log concave.*

Proof. A discrete distribution is log concave if and only if it exhibits an increasing failure rate (Lai & Xie, 2006). $MRL_{it}(q)$ is nonincreasing for discrete distributions exhibiting an increasing failure rate (Lai & Xie, 2006). *Q.E.D.*

With the lemmas in hand, we can proceed to the proofs of the propositions from the text:

Proof of Proposition 1. Substituting from equation (1) into $\Delta TC_{it} = TC^o_{it} - TC^p_{it}$ and rearranging yields

$$\begin{aligned}\Delta TC_{it} &= N_t\left[l^o_{it}\overline{F}_{it}(a^o_{it}) - l^p_{it}\overline{F}_{it}(a^p_{it})\right] \\ &\geq N_t l^p_{it}\left[\overline{F}_{it}(a^o_{it}) - \overline{F}_{it}(a^p_{it})\right].\end{aligned}$$

The second line holds since $l^o_{it} \geq l^p_{it}$. The second line is nonnegative since $a^o_{it} \leq a^p_{it}$, implying $\overline{F}_{it}(a^o_{it}) \geq \overline{F}_{it}(a^p_{it})$ because $\overline{F}_{it}$ is nonincreasing by lemma 1.

Substituting from equation (2) into $\Delta TW_{it} = TW^o_{it} - TW^p_{it}$ and rearranging implies that ΔTW_{it} equals

$$N_t\left[l^o_{it}\sum_{q>a^o_{it}}(q-a^o_{it})f_{it}(q) - l^p_{it}\sum_{q>a^p_{it}}(q-a^p_{it})f_{it}(q)\right]$$

$$\geq N_t l^p_{it}\left[\sum_{q>a^o_{it}}(q-a^o_{it})f_{it}(q) - \sum_{q>a^p_{it}}(q-a^p_{it})f_{it}(q)\right]$$

$$= N_t l^p_{it}\left[\sum_{q>a^p_{it}}(a^p_{it}-a^o_{it})f_{it}(q) + \sum_{q\in(a^o_{it},a^p_{it}]}(q-a^o_{it})f_{it}(q)\right].$$

The second line again holds since $l^o_{it} \geq l^p_{it}$. The last line follows from rearranging terms. It is easy to see that both terms in brackets are nonnegative and hence the whole expression is nonnegative. *Q.E.D.*

Proof of Proposition 2. Comparing equations (2) and (A1) shows

$$\begin{aligned}TW^p_{it} &= N_t l^p_{it}\overline{F}_{it}(a^p_{it})MRL_{it}(a^p_{it}) \\ &= TC^p_{it}MRL_{it}(a^p_{it}),\end{aligned}$$

where the second line follows from substituting from equation (1). Similarly, $TW_{it}^{o} = TC_{it}^{o} MRL_{it}\left(a_{it}^{o}\right)$. Thus,

$$\begin{aligned}
&\frac{\Delta TW_{it}}{TW_{it}^{p}} - \frac{\Delta TC_{it}}{TC_{it}^{p}} \\
&= \frac{TC_{it}^{o} MRL_{it}\left(a_{it}^{o}\right) - TC_{it}^{p} MRL_{it}\left(a_{it}^{p}\right)}{TC_{it}^{p} MRL_{it}\left(a_{it}^{p}\right)} - \frac{TC_{it}^{o} - TC_{it}^{p}}{TC_{it}^{p}} \\
&= \frac{TC_{it}^{o}\left[MRL_{it}\left(a_{it}^{o}\right) - MRL_{it}\left(a_{it}^{p}\right)\right]}{TC_{it}^{p} MRL_{it}\left(a_{it}^{p}\right)}.
\end{aligned}$$

The sign of the last expression is determined by the bracketed factor. The log concavity of $f_{it}(q)$ together with $a_{it}^{o} \leq a_{it}^{p}$ imply that $MRL_{it}\left(a_{it}^{o}\right) \geq MRL_{it}\left(a_{it}^{p}\right)$ by lemma 2. Hence the last expression is nonnegative. Thus, $\Delta TW_{it}/TW_{it}^{p} \geq \Delta TC_{it}/TC_{it}^{p}$. *Q.E.D.*

Proof of Proposition 3. We examine the simultaneous effect of an improvement in the availability of article i captured by a weak increase in l_{nit} and a weak decrease in a_{nit} for all $n = 1, \ldots, N_t$ on expected own and rival cites. First, we examine the effect on $E\left(TC_{it}^{*} | a_t, l_t, q_t\right)$. We have

$$E\left(TC_{it}^{*} | a_t, l_t, q_t\right) = \sum_{n=1}^{N_t} l_{nit} E\left(C_{nit}^{*} | a_{nt}, q_{nt}, L_{nit} = 1\right). \tag{A2}$$

The expectation on the left-hand side is taken with respect to the distribution of L_t and on the right-hand side with respect to the distribution of L_{nt}^{i}, where L_{nt}^{i} is formed by starting with vector L_{nt} and substituting a 1 for the component for article i. In other words, L_{nt}^{i} delineates a set of articles of which n is aware when n is certainly aware of i. Consider one of the terms in the sum on the right-hand side of equation (A2), say,

$$l_{nit} E\left(C_{nit}^{*} | a_{nt}, q_{nt}, L_{nit} = 1\right). \tag{A3}$$

A change in l_{nit} affects equation (A3) only through the leading factor. The expectation is independent of l_{nit} because it is conditioned on a realized value, $L_{nit} = 1$, of the Bernoulli random variable of which l_{nit} characterizes the distribution. Clearly, then, a weak increase in l_{nit} will result in a weak increase in equation (A3). A change in a_{nit} affects equation (A3) through its effect on the conditional expectation. A closer look at this conditional expectation gives

$$E\left(C_{nit}^{*} | a_{nt}, q_{nt}, L_{nit} = 1\right) = \sum_{L_{nt}^{i} \in \{0,1\}^{I_t}} Pr\left(L_{nt}^{i}\right) C_{nit}^{*}\left(q_{nt}, a_{nt}, L_{nt}^{i}\right), \tag{A4}$$

where

$$Pr\left(L_{nt}^{i}\right) = \prod_{j \neq i} \left[l_{njt} L_{njt} + \left(1 - l_{njt}\right)\left(1 - L_{njt}\right)\right]$$

is a multivariate Bernoulli probability function. Note that we have explicitly written the arguments of C_{nit}^{*} in equation (A4) to emphasize the dependence of this maximizer on them. In the last of its arguments, we can restrict attention to vectors with a 1 in the component for article i because the optimal citation decision for an article of which n is not aware is irrelevant. We can rewrite the objective function (3) of which $C_{nit}^{*}\left(q_{nt}, a_{nt}, L_{nt}^{i}\right)$ is a component of the vector of maximizers as

$$U_n = \left[-a_{nit} C_{nit}\right] + \left[B_n\left(L_{nt} \cdot C_{nt} \cdot q_{nt}\right) - \sum_{j \neq i} L_{njt} C_{njt} a_{njt}\right],$$

where the brackets separate the part of U_n that depends on a_{nit} from the part that does not. The first bracketed expression is supermodular in C_{nit} and $-a_{nit}$ by theorem 6 of Milgrom and Shannon (1994) because the cross-partial derivative is positive. Thus, by corollary 4 of Milgrom and Shannon (1994), C_{nit}^{*} is nondecreasing in $-a_{nit}$ and therefore nonincreasing in a_{nit}.

We can combine the individual results, which all point in the same direction, moving in reverse from equation (A4) to (A3) to (A2) to show that $E\left(TC_{it}^{*} | a_t, l_t, q_t\right)$ weakly increases with a simultaneous weak increase in l_{nit} and weak decrease in a_{nit} for all $n = 1, \ldots, N_t$. This establishes the own-citation result.

The initial steps used to establish the rival-citation result are similar to those above. The last step involves the application of theorem 5 from Milgrom and Shannon (1994) with U_n as the objective function, C_{n1t} and $1 - C_{n2t}$ as the endogenous variables, and $-a_{n1t}$ as the exogenous variable. The theorem states that C_{n1t}^{*} and $1 - C_{n2t}^{*}$ are nondecreasing in $-a_{n1t}$ if:

a. U_n is supermodular in C_{n1t}^{*} and $1 - C_{n2t}^{*}$.
b. U_n exhibits increasing differences in C_{n1t}^{*} and $-a_{n1t}$.
c. U_n exhibits increasing differences in $1 - C_{n2t}^{*}$ and $-a_{n1t}$.

By Milgrom and Shannon (1994, theorem 6), all three properties follow if the associated cross-partial derivatives are nonnegative. Differentiating,

$$\frac{\partial^2 U_n}{\partial C_{n1t} \partial\left(1 - C_{n2t}\right)} = -L_{n1t} L_{n2t} q_{n1t} q_{n2t} B_n'' \tag{A5}$$

$$\frac{\partial^2 U_n}{\partial C_{n1t} \partial\left(-a_{n1t}\right)} = L_{n1t} \tag{A6}$$

$$\frac{\partial^2 U_n}{\partial\left(1 - C_{n2t}\right) \partial\left(-a_{n1t}\right)} = 0. \tag{A7}$$

The concavity of B_n implies that equation (A5) is nonnegative, establishing property (a). Equations (A6) and (A7) are obviously nonnegative, establishing properties b and c, respectively.

This proves that C_{n1t}^{*} and $1 - C_{n2t}^{*}$ are nondecreasing in $-a_{n1t}$. Thus, a weak decrease in a_{n1t} causes both a weak increase in C_{n1t}^{*} and a weak decrease in C_{n2t}^{*}. The rest of the proof is then similar to that for the own-citation result. *Q.E.D.*