Did Online Access to Journals Change the Economics Literature?

Mark J. McCabe
School of Information, University of Michigan, 105 S. State Street, Ann Arbor, MI 48104; email: mccabe@umich.edu

Christopher M. Snyder
Department of Economics, Dartmouth College, 301 Rockefeller Hall, Hanover, NH 03755; email: chris.snyder@dartmouth.edu

September 2011

Abstract: Does online access boost citations? The answer has implications for issues ranging from the value of a citation to the sustainability of open-access journals. Using panel data on 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 we add fixed effects as controls. The absence of an aggregate effect masks heterogeneity across platforms: JSTOR boosts citations around 10%; ScienceDirect has no effect. We examine other sources of heterogeneity including whether JSTOR benefits "long-tail" or "superstar" articles more.

JEL Codes: L17 (open source products and networks), O33 (technological change: diffusion processes), A11 (role of economists)

Acknowledgements: The authors are 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, 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, 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). The authors are grateful to Mark Bard, Jamie Bergeson-Bradshaw, Yilan Hu, Ella Kim, Scot Parsley, 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. This research was generously supported by a grant from the Andrew W. Mellon Foundation.

1. Introduction

Could an economist quadruple his or her citation count just by publishing in an online journal rather than one available only in print? Our initial interest in this question was prompted by the huge effects of online access found in previous empirical studies. For example, in the study of a cross section of medical journals over the period 1995-2000, Curti et al. (2001) found online journals generated 54% more cites per article than print-only journals. Lawrence (2001) studied a sample of computer-science conference proceedings that exhibited within-proceedings variation in access, with some articles made available online and the rest only in print. In the average proceedings, online articles received 336% more cites than print.

It would not be surprising if convenient online access to the full text of an article boosts its citations.[1] 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 almost implausibly large results for the citation effect in these previous studies suggests they are biased upward. A likely 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, in Lawrence (2001), there is no mention that 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 separately identifying the effect of online access from unobserved quality. However, 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.[3] If a small change in the convenience of access can cause a quadrupling of citations, then the typical citation may be of

[1] Digital bibliographic data for articles was provided to libraries in the 1970s (Lancaster and 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 and Neufang 1995). Around the same time, academic journals began providing Internet access to the full text some articles.

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

[3] We follow the traditional approach in considering the raw citation count as the "currency." See, e.g., Palacios-Huerta and Volij (2004) and Bollen et al. (2005) for a discussion of alternative impact metrics.

marginal value, used to pad the reference section of citing articles 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 and Snyder 2005, 2007, 2010; Jeon and Rochet 2010) uses a two-sided-market model[4] to assess which pricing scheme would dominate in equilibrium and which would generate the most social surplus. The answer to both questions hinges on the elasticities of demand on the author and reader sides. How much more an author would pay for better access by readers to his or her article depends on how this access translates into readership and citations. If online access quadruples 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 that this open-access equilibrium have desirable efficiency properties.[5] On the other hand, if the citation benefit is low, author demand may be so elastic that open access is unsustainable in equilibrium and/or socially inefficient. Previous research on the effect of open access on citations find the same huge results as the literature on online access cited above[6] and likely suffer from the same biases,[7] 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.

[4] 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 exert 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.

[5] 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.

[6] 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.

[7] The decision by an author to self-archive (studied by Harnad and 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.

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 in an earlier period and Google Scholar more recently.[8] 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 and Dosi 1998).

We address the econometric challenge of separating the citation effect of online access from unobserved quality effects by assembling a large panel dataset, described in Section 3, on the citations indexed by Thomson ISI in the 1980-2005 period to all the articles published since 1956 in a sample of the top 100 economics and business journals. We merge in hand-collected data on the date that each journal volume was made available on the Internet and via which platforms (i.e., the journal's own website, JSTOR, or several other major Internet platforms). The panel nature of the dataset allows us to control for unobserved quality effects by including fixed effects for journal volumes.[9] 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 (only after an embargo window, for example). As discussed in the literature review in Section 2, 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.

Section 4 outlines the econometric specification, informed by a theoretical model provided in the Appendix. Section 5 presents the results. The first set of results show that the same huge effects of online access found in the previous literature can be generated if fixed

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

[9] Although we have article-level data, because the date of online access does not vary across articles within a volume in our data, our analysis is conducted at the level of the journal volume. Hence the most refined set of fixed effects for content are journal-volume fixed effects.

effects capturing the quality level of journal volumes are omitted. Once appropriate fixed effects are included, however, the aggregate result cannot be distinguished from zero. 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 policy makers that facilitating access disproportionately benefits authors in developing countries where library resources may otherwise be limited.

Section 6 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 clothing (Brynjolfsson, Hu, and Simester 2007) to video sales (Elberse and 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.

2. 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 controlling for quality by using panel data on 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 impart a positive bias to the variables of interest because online access also varies systematically across age and time. We demonstrate the point

concretely in Table 2, where for example we reproduce a similar estimate to the 26% for economics and business in Evans and Reimer (2009), but then show this estimate disappears when the full suite of age and time effects is added.[10]

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 increase in the number of journals available via JSTOR to the authors of an article increases cites to JSTOR journals and decreases cites to non-JSTOR journals. Absent fixed effects for citing authors, the result could be an artifact of heterogeneity in citing patterns across different institutions. For example, authors at higher-ranked institutions (which generally have more JSTOR access) may have a greater propensity to cite leading journals (the journals carried by JSTOR tend to be higher ranked).[11]

Two recent 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 upon publication, the rest receiving the usual treatment of restricted/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 (2009) 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

[10] 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 a electronic database, Fulltext Sources Online, for information on online histories for journal 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 (1995-98), contains inaccuracies in online-access dates, and omits important channels (including JSTOR).

[11] This problem was addressed to some extent by Parker, Bauer, and Sullenger (2003) and De Groote, Shultz, and Doranski (2006). They focus on 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 allows for author fixed effects, 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.

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 "non results" from these two studies are consistent with our finding of no aggregate effect.[12] 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 of the field, only made 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.

3. Data

Our analysis is based on a sample of 100 journals in economics and business.[13] 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

[12] 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 open versus fee access for an online journal.

[13] We followed the subject categorization of our data vendor, Thomson ISI, focusing on economics and business because of our interest in the journal market in our own discipline and knowledge of its institutional details. We conducted a similar analysis—not reported here for space considerations—of 100 journals in each of history and biomedicine. Journal articles in history generate remarkably few subsequent citations in journals, possibly reflecting history's reliance on books rather than journals, leading to noisy results. The results from biomedicine are similar in many respects to economics and business, providing some confidence in the latter's representativeness.

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 1985-2004 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.[14] Overall, the sample includes 63 journals in economics and 37 in business. The sample has a extensive representation of digital channels besides JSTOR. A variety of online channels are represented in addition to JSTOR. For example, 30 journals published by Elsevier are eventually available online via ScienceDirect.

The dataset merges citations data together with historical information on online availability. The citations data was 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 article and the cited article. 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.[15] The major aggregators considered 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 dataset 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

[14] 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, the *Canadian Journal of Economics* and the *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 later 89 among economics journals.

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

level—the volume—comprising all of the articles a journal publishes in a given 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 dataset 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 dataset'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 dataset. All journals were founded by 1988, the earliest, the *Journal of Institutional and Theoretical Economics*, 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 *Econometrica*). The heterogeneity in number of citations which 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 lifespan of the average journal volume.[16,17] Citations peak in the fifth year

[16] 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 4 provides a more formal discussion of the estimation procedure, due to Wooldridge (1999).

[17] 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

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 3/4 of the cites received in the initial year of publication. The 95% confidence interval shows that the estimates start out very 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 over time, reaching a level by the end of the sample that is 120% higher than in the base year of 1980. This increase in citations is due to a number of factors including the increase in the number of citing journals indexed by ISI, an increase in articles per journal, and an increase in the number of references per citing article.

The last row of 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.

4. Empirical Methodology

4.1. 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:

$$E(CIT_{vt}|\alpha_v, x_{vt}) = \exp(\alpha_v + x_{vt}\beta), \quad (1)$$

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

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 4.3 for further discussion.

analysis in the previous section was conducted at the article level, the econometric analysis will use volume-level observations because, as noted above, 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).[18]

Including the volume fixed effects α_v in (1) 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, as discussed in the Appendix, "quality" 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:

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

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 which are allowed to vary not just across journals but across five blocks of ten publication years for each journal.[19] 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

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

[19] 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.

full online access, that is, availability of the entire volume's content for the entire year, but also include controls for partial online access.

4.2. Heterogeneity in Online Access

Intuition suggests (and the theoretical analysis in the Appendix bears out) that the online-access effect may exhibit substantial heterogeneity, 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 6 to this question.

4.3. Identification Strategy and Challenges

Footnote 17 discussed the "identification problem," i.e., 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 only 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.[20]

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 were 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

[20] Our primary regressions include a richer set of fixed effects than was used to construct the figures referenced by footnote 17, 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.

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 (1)]. 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 choose when to place different volumes online based on their relative numbers of citations. The example of the *American Economic Review* (*AER*), shown in Figure 5, suggests that the online-access variable is plausibly exogenous. 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 only available 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 the volumes' relative citations.

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, 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 (1956-89) of its backfiles online. More recent volumes were added in a stream, but 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. Paradoxically, this bunching of online availability provides a useful source of variation because simultaneous access affects different volumes at different

points in their age profiles. For example, JSTOR's initial tranche of over 30 volumes of the *AER* is a shock to the 1956 volume in its fortieth year but the 1989 volume in its sixth. The less pronounced but still evident bunching for more recent content provides a similarly useful source of variation. Online availability varies across journals as well as volumes within a journal. As shown in Figure 5, JSTOR added the big tranche of *AER* backfiles in 1996. While some other journals were added to JSTOR at the same time, others were added at various points later (for example the *Economic Journal* in 1998 and the *Review of Economic Studies* in 1999).

5. Results

5.1. Alternative Specifications

Table 2 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.[21] 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 are detailed in the notes for the table. Regression coefficients have been converted into a form interpretable as proportionate increases: a zero result corresponds to no measured effect from online access; a negative result corresponds to online access causing a reduction in cites; 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 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 astonishingly high online-access effects found in these studies. For example, the first coefficient of 5.195 has the interpretation that volumes published in 1956-65

[21] 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 (1995-99) and later ones (2000-05). We use shaded boxes as a device to indicate which results are coming from the same and which from separate regressions.

received a more than a 500% fold boost in citations from online access in the years 1995-99 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. The results are again 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 zero 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.[22] While our underlying data source is different (see footnote 10), 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 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 (2000-05) 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 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 5 but adds publication-year × citation-year fixed

[22] 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 that 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).

effects to control for secular trends.[23] The digitization effect for the 2000-05 period disappears. This suggests that Evans and Reimer's (2009) results are spurious, generated by omitted-variable bias. Omitting time effects forces the open-access variable to pick up age effects (which are positive immediately to the right of the peak in Figure 1) and secular trends in citations (which are positive over the whole range in Figure 2). 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.

5.2. Individual Channels

As discussed in Section 4 (see also the Appendix), 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.

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.[24] For JSTOR, we have data on the number of institutions subscribing to

[23] To be consistent with our preferred specification in column 4, we also add a quadratic age profile for each block of 10 volumes of a journal, but this inclusion changes the results from column 5 less than the inclusion of publication-year × citation-year fixed effects.

[24] We also ran an indicator specification consistent with the aggregate regressions in Table 2. That is, separate indicators were included for online access through the channel in the early (1995-99) and late online periods (2000-05). 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.

different packages of its online journals. For Elsevier, we have subscriptions for journal backfiles (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 just to its backfiles.[25] Figure 7 graphs trends in JSTOR and Elsevier backfile subscriptions. During 1995-97, although JSTOR had content posted online, it had hardly any institutional subscribers. After 1997, JSTOR subscriptions increased linearly with time throughout the rest of the sample. Subscriptions to Elsevier backfiles increased fairly linearly with time after 2001, but the slower rate and later start kept Elsevier backfiles to only about a quarter of JSTOR subscriptions in 2005.

Table 3 reports the results for JSTOR and Elsevier backfiles subscriptions. For ease of interpretation, the parameters have been converted into elasticities. For example, 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 1956-65 content). A pair of estimates are provided for JSTOR: one for cases in which JSTOR is the sole channel and another for cases in which JSTOR duplicates the access via some other channel. As 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 backfiles. 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 falls fairly consistently as one moves down column 1 from the oldest publication-year block (1956-65) to the most recent. When JSTOR provides access along with some other channel (column 2), this set of impacts is similar, ranging between 1.7% and 18.6%. The fact that access to other channel does not seem to impair the marginal contribution of JSTOR access suggest that other channels are not good substitutes for

[25] 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.

JSTOR. In contrast to the JSTOR results, the Elsevier elasticities are small, sometimes negative, and statistically insignificant. Thus, online access to Elsevier's backfile content through its own website (ScienceDirect) appears to provide no citation boost; JSTOR appears to have a uniquely strong effect on citations, an effect which is generally strongest for the oldest content.

5.3. 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 policy makers that online access will have a disproportionate benefit for scholars in developing countries because their libraries have the smallest print collections.

There are a number of alternatives for defining the location of an article with several authors; for simplicity we use the location of the first citing author. Articles with U.S.-based first authors are responsible for a majority (60%) of cites, reflecting the influence of that region in the fields of economics and business. The English-speaking West is responsible for 18% of cites, non-English speaking West for 16%, and the rest of the world for the remaining 6%.

Table 4 presents the regression results for the breakdown by region. The reported results will again be for the two most important individual channels, JSTOR and Elsevier. 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 3; the only differences are that the left-hand side variable is the number of citations from 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 United States, the U.S. results 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 3. Elsevier continues to show no significantly positive online access effect. Indeed, the negative elasticity (-0.104) for Elsevier content published in 1976-85 is significant at the 10% level. The pattern of results is also similar for the other regions except for the non-English speaking West. In that region, JSTOR has no statistically significant positive effect. The only statistically significant

results are the negative ones for the most recent content (1999-2005). The result does not seem to be due to the relative lack of 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 et al. (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 journals ranked by Lubrano et al. (2003) according to an objective citations measure.

Overall, the regional breakdown does not support the conclusion that developing countries benefit disproportionately from online access. The rest of the world does show a benefit from online access but not different in kind from the United States and the English-speaking West.

We performed other breakdowns of the result along other possible sources of heterogeneity. One other possible source was 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 proportion of cites in our data accounted for by first citing authors at that institution. The distribution of citations showed the expected skewness with the top 100 citing institutions responsible for 73% of citations in our data. 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 vs. 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%.

6. 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, Hu, and Simester 2007) to video sales (Elberse and Oberholzer-Gee 2008). As discussed in the Appendix, in theory, 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 of 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 those articles that are cited, suggesting a superstar effect. Contrasting Evans' findings, we will show that, in the case of JSTOR, the effect of online access is fairly uniform across the distribution of articles, benefiting superstar and more obscure articles alike.

In other words, the typical power law relationship between ranked articles and citation counts is shifted up but its shape is not changed. In direct contrast to Evans (2008), we find that online access increases the fraction of articles receiving any cites. The discrepancy likely lies in methodological problems in Evans (2008) discussed above.

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 then estimate the online-access effect in later years using a 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 receiving any cites. The first approach is described in Section 6.1 and the second in Section 6.2.

6.1. 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 and 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 due to selection of the sample based on residuals, 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 re-aggregate 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 (1) 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 only using citation years $t_v >$

$t_{vw} + t_{vg}$ for volume v.[26]

The results are reported in Table 5. The specification is identical to that in Table 3, the only differences being the ones just mentioned, that the aggregated results in Table 3 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.[27]

The results for the highest-citation (80-100) quintile in columns 9 and 10 are very similar to the corresponding aggregate results in Table 3 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 out of the five publication-year blocks. Similar 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 1-8, 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

[26] 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 per-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.

[27] Online access to Elsevier backfiles produces no statistically significant citation effects even when broken down at the quintile level, reinforcing the conclusions about Elsevier from Table 3.

positive JSTOR effects can be observed throughout the distribution of articles, from less popular ones in the long tail to the superstars.

6.2. 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). The PFP estimator assumes a conditional mean of the following general form

$$E(FCIT_{vt}|\alpha_v, x_{vt}) = \Phi\left(\alpha_v + x_{vt}\beta + \bar{x}_{j(v)t}\xi\right) \quad (3)$$

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, $\bar{x}_{j(v)t}$ is the mean value of regressors x_{vt} across volumes for the same journal, and β and ξ are parameter vectors.[28] The reported results Table 6 are average partial effects, obtained by differentiating (3) with respect to the variable of interest, online-channel subscriptions, and then summing the result across observations in the sample. Average partial effects are converted into elasticities for comparison to results from previous methods.

Table 6 again reports the results just for the important online channels with subscription data: JSTOR and Elsevier backfiles. Although involving a different left-hand-side variable than in Table 3—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

[28] The estimator can be implemented in Stata by regressing $FCIT_{vt}$ on a constant, regressors x_{vt}, and regressor means $\bar{x}_{j(v)t}$, 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.

JSTOR subscriptions increases the fraction of cited articles by 17.6 percentage points for the earliest content (1956-65 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 backfiles has no measurable effect on the fraction of cited articles.

Overall, the results from Table 6 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 subsection 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 backfiles at any point in the ranking of articles by citations.

7. Conclusions

Our results for the effect of online access on citations to economics articles 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 huge levels found in the previous literature down to a precisely estimated value of zero. 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 recent studies (e.g., Evans and Reimer 2009) which 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 dataset 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 backfile history up to the embargo window, and it was an early entrant in the market. Indeed, JSTOR offered online access to backfiles five years before ScienceDirect, a long time for users to learn to use the platform and share their experience with colleagues.[29] 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 Appendix model of article search and acquisition: 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 backfiles 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.[30]

We disaggregated the results in other dimensions, focusing mainly on JSTOR because this is the case most likely to give nontrivial results. The JSTOR results were surprisingly uniform across these other breakdowns. Based on our finding that older content received the greatest proportional citation boost from online access, we expected that the effects would be larger for categories that might be viewed as suffering a disadvantage which online access might help overcome. For example, 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

[29] 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.

[30] 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 and Stern 2011).

effect for them. Likewise, authors from lower-ranked institutions or from countries outside of core for economics publishing might show a disproportionate increase in citations from online access. Instead we found that the citation benefit was fairly uniform across cited journals by rank, and across citing authors by rank of citing institutions or region. The one anomaly was the effect of JSTOR on citing authors in non-English-speaking Europe. 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. We find no evidence supporting the claim that citing authors in developing countries would receive a disproportionate benefit from more convenient access than, say, those in the United States.

We further disaggregated the results by binning the articles into quintiles based on citation rank in a pre-period. 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 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.[31]

Tying the results back to the broader policy issues considered in the introduction, 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

[31] Hervas-Drane (2010) provides a model in which long-tail and superstar effects operate simultaneously in retail markets.

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 and Snyder 2005, 2007, 2010; Jeon and 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. Whether the social value in terms of scholarly productivity—and the spillover benefits for economic productivity more broadly—exceed the cost of constructing and maintaining the channel are beyond the scope of this paper, but at least a plausible case can be made. The greater impact of JSTOR relative to other channels does support the utility of some of JSTOR's attributes, including its stability and its coverage of a large number of journals and a complete set of backfiles for most of these. While they may not revolutionize the scholarly literature, next-generation technologies such as Google Scholar advance some of these same desirable features and thus promise to continue making measurable contributions to scholarly productivity.

Appendix: Theoretical Model

This appendix provides a theoretical model motivating the empirical specification used in Section 4.

We model citations by first characterizing the behavior of a representative citing author n and then aggregating over $n = 1, \ldots, N_t$ where N_t is the total number of citing authors who publish an article in year t (each citing author is assumed to publish one article). Let $i = 1, \ldots, I_t$ index economics articles that can be potentially cited in year t, where I_t is the total number of them. A cite by author n to article i results from a two stage process: first, the article and its value to one's project is identified through search; second, a full-text copy of the article is acquired and the information that will make up the cited passage extracted.

A.1. Search Stage

To characterize the search stage, let s_n be author n's expenditure of effort in searching through the literature. This search yields a probability $\pi_{ni} \in [0,1]$ that the author finds article i. Conditional on finding the article, the author learns its match quality for him or her, Q_{ni}. 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 as topic).

The probability that author n finds article i is

$$\pi_{ni} = \pi(s_n, M_t, Q_{ni}, OA_{it} l_{ni} c_{it}). \tag{4}$$

Assume π is increasing in its first argument, s_n, implying that an article is more likely to be found the harder the citing author searches the literature. The second argument, M_t, is a vector of overall "market" conditions, including the size of the pool of citable articles I_t (since π is the probability of finding a given "tree" in a "forest" of articles, it is plausibly declining in the size of the "forest"), the technology available for searching through these other articles, and so forth. Assume π is weakly increasing in its third argument, Q_{ni}, implying that the citing author can more easily find articles with good match quality. Taking an extreme case, authors may be well acquainted with the handful of seminal articles in their research areas without having to search at all, in which case $\pi_{ni} = 1$ even if $s_n = 0$ for these. The last argument, which is the product of three factors, captures the contribution of online access to search. The first factor, OA_{it}, is the indicator for online access to the channel under consideration (or a vector of indicators if all channels are being considered jointly); the second factor, l_{ni}, is an indicator for whether author n's library subscribes to the channel or channels providing online access; and the third, c_{it}, is a vector of characteristics of the channel or channels. The attributes of the online channel will affect the ease of search: browsing is easier the more journals a channel covers and the more complete set of volumes for each journal.

A.2. Acquisition Stage

In the second stage, citing author n decides of which identified articles to acquire full-text copies, comparing the benefit of acquisition, Q_{ni}, to the cost, a_{ni}. While both benefit and cost are known to n at the time the decision is made in the second stage, in the first stage a_{ni} is

uncertain, a continuous random variable on $[0, \infty)$ with distribution function $F(a_{ni}, OA_{it} l_{ni} c_{it})$. The second argument allows the acquisition cost to depend on the same online-access factors as the search process in the first stage. Assume that better online availability (in the sense that the indicators OA_{it} and l_{ni} are turned on or the characteristics c_{it} of the online channel are more favorable) reduces the acquisition cost in the sense of first-order stochastic dominance. Author n acquires article i (and automatically goes on to cite it) if $Q_{ni} > a_{ni}$. Combining the probability that the article is worth acquiring with the probability article i was identified in the first-stage search yields the following overall probability of citation from author n to article i:

$$\pi_{ni} F(Q_{ni}, OA_{it} l_{ni} c_{it}). \tag{5}$$

By equation (4), the leading factor, π_{ni}, in (5) is a function of the endogenous search effort, s_n. Author n chooses s_n to maximize the benefit minus the cost of search:

$$\int_0^{I_t} \pi(s_n, M_t, Q_{ni}, OA_{it} l_{ni} c_{it}) \left[\int_0^{Q_{ni}} (Q_{ni} - a) dF(a, OA_{it} l_{ni} c_{it}) \right] di - s_n. \tag{6}$$

The outer integral in (6) integrates over the population of citable articles at time t. Each article is taken to be infinitesimal in the sense that the characteristics of any one do not materially affect the author's search decision. The inner integral is the expected return of acquiring an identified article which passes author n's benefit-cost test.

The optimal search effort $s_n^*(M_t)$ can be written as a function of market conditions alone because the characteristics (Q_{ni}, OA_{it}, l_{ni}, c_{it}) of individual, infinitesimal articles integrate out. Only the distribution of these characteristics across articles matters; and this distribution can be considered a component of M_t. Substituting $s_n^*(M_t)$ into the expression for π_{ni} in equation (4) and then this value of π_{ni} into (5) gives a single author n's probability of citing article i. Summing over authors $n = 1, \ldots, N_t$ yields total citations to article i at time t:

$$CIT_{it} = C\left(M_t, \{Q_{ni}\}_{n=1}^{N_t}, OA_{it} \xi\left(\{l_{ni}\}_{n=1}^{N_t}, c_{it}\right)\right). \tag{7}$$

According to equation (7), the number of citations is a function of market factors M_t, the distribution of the quality of the match with article i in population of citing authors $\{Q_{ni}\}_{n=1}^{N_t}$, and the nature of online access. The nature of online access is given by the product of the online-access indicator OA_{it} and a function ξ of the distribution across the population of citing authors of library subscriptions to the online channel or channels $\{l_{ni}\}_{n=1}^{N_t}$ and the characteristics of the online channel or channels c_{it}.

A.3. Comparative Statics

Equation (7) has straightforward comparative-statics properties. The effect of market forces M_t will depend on which of its components is considered. An increase in the population of citing authors will tend to increase cites; an increase in the population of citable articles will tend to reduce cites to a given article by diluting search effort; a change in professional norms toward including a larger literature review and bibliography in each article will tend to increase cites.

An upward shift in the distribution of match qualities $\{Q_{ni}\}_{n=1}^{N_t}$ will generate more successful searches and induce more acquisition effort for identified articles, both effects tending to increase cites. Turning to the last argument of C, we expect online access to increase cites by making search and acquisition effort more productive for a given article. The online-access effect is mediated through several other layers—the citer's library and the channels to which this library subscribes—and the nature of these other layers will contribute to the magnitude of the online-access effect. The more libraries that subscribe to online channels and the better the attributes of the online channels, the more online access will boost cites. However, it is theoretically possible that online access could end up reducing cites if the publisher uses the occasion to raise prices enough to cause a large drop in library subscriptions.

Interactions between OA_{it} and the first two arguments of C may lead to heterogeneous online-access effects. The effect of online access may gradually increase as citing authors grow familiar with the use of the Internet in their literature search. The online-access effect may vary with the quality of the article. Seminal articles may be so valuable that they are identified and acquired even if the procedure is fairly costly, so that online access would have little effect for them, corresponding to a negative cross partial between the second and third arguments of C. Arguments could also be proposed for the cross partial to have the opposite sign.

A.4. Empirical Implications

We conclude the Appendix by highlighting where the intuitive claims made in the text about the appropriate econometric specification show up in the model. First, consider the concern the online access estimate may be biased upward if quality is not controlled for. This arises in the model if the second and third arguments of C in equation (7) are positively correlated but a proxy for the second argument is not included in the regression. Second, consider the concern that adequate age controls need to be included for consistency. This arises in the model recognizing that an important factor in match quality is article age. Given that direct controls for the distribution of match quality $\{Q_{ni}\}_{n=1}^{N_t}$ are impossible to include in the regression and that the fixed volume effects α_v only account for age-invariant sources of match quality, omission of a rich set of age controls can bias the results for online access because they can spuriously reflect age-varying sources of match quality. Third, the results of interest may be biased if adequate controls for time effects are not included. In the model, a positive correlation between the first and third argument of C would generate such a bias if M_t is not fully controlled for. Fourth, heterogeneity in the online-access effect may arise from interactions between the last argument of C and the others. It may also stem from the multiplication of OA_{it} with ξ in the last argument because ξ is itself a function of channel attributes such as number of subscribers and other factors. Finally, the model is consistent with both "long-tail" and "superstar" effects, depending on the sign of the cross partial between the second and third arguments of C, a negative cross partial leading to a "long-tail" effect and a positive one leading to a "superstar" effect.

References

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

Armstrong, Mark. (2006) "Competition in Two-Sided Markets," *Rand Journal of Economics* 37: 668-91.

Bergstrom, Theodore. (2001) "Free Labor for Costly Journals?" *Journal of Economic Perspectives* 15: 454-74

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

Bollen, Johan, et al. (2005) "Toward Alternative Metrics of Journal Impact: A Comparison of Download and Citation Data," *Information Processing & Management* 41: 1419-40.

Brynjolfsson, Erik, Yu (Jeffrey) Hu, and Duncan Simester. (2007) "Goodbye Pareto Principle, Hello Long Tail: The Effect of Search Costs on the Concentration of Product Sales," MIT Sloan School working paper.

Craig, Iain D., et al. (2007) "Do Open Access Articles Have Greater Citation Impact? A Critical Review of the Literature," *Journal of Informetrics* 1: 239-48.

Curti, Moreno, et al. (2001) "Impact Factor and Electronic Versions of Biomedical Scientific Journals," *Haematologica* 86:1015-20.

Davis, Philip M., et al. (2008) "Open Access Publishing, Article Downloads, and Citations: Randomised Controlled Trial," *British Medical Journal* 337: 568-73.

De Groote, Sandra L., Mary Shultz, and Marceline Doranski. (2005) "Online Journals' Impact on the Citation Patterns of Medical Faculty," *Journal of the Medical Library Association* 93: 223-28.

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

Dewatripont, Mathias, et al. (2006) *Study on the Economic and Technical Evolution of the Scientific Publication markets in Europe.* Brussels: European Commission Directorate General for Research.

Dosi, Giovanni. (1988) **"**Sources, Procedures, and Microeconomic Effects of Innovation," *Journal of Economic Literature*, 26: 1120-71.

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

Elberse, Anita and Felix Oberholzer-Gee. (2008) "Superstars and Underdogs: An Examination of The Long Tail Phenomenon in Video Sales," Harvard Business School working paper no. 07-015.

Evans, James. (2008) "Electronic Publication and the Narrowing of Science and Scholarship," *Science* 321: 395-99.

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

Eysenbach, Gunther. (2006) "Citation Advantage of Open Access Articles," *PLoS Biology* 4: 692-98.

Fauli-Oller, Ramon and Joel Sandonis. (2010) "On the Profitability and Welfare Effects of Downstream Mergers," University of Alicante working paper.

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

Furman, Jeffrey L. and Scott Stern. (2011) "Climbing Atop the Shoulders of Giants: The Impact of Institutions on Cumulative Research," *American Economic Review* 101: 1933-63.

Gaule, Patrick and Nicholas Maystre. (2009) "Getting Cited: Does Open Access Help?" University of Geneva working paper, SSRN abstract no. 1427763.

Harley, Diane, et al. (2010) *Assessing the Future Landscape of Scholarly Communication: An Exploration of Faculty Values and Needs in Seven Disciplines*. UC Berkeley: Center for Studies in Higher Education. Retrieved from http://escholarship.org/uc/cshe_fsc.

Harnad, Steven and Tim Brody. (2004) "Comparing the Impact of Open Access (OA) vs. Non-OA Articles in the Same Journals," *D-Lib Magazine*, 10 no. 6.

Hervas-Drane, Andres. (2009) "Word of Mouth and Taste Matching: A Theory of the Long Tail" NET Institute working paper no. 07-41.

Inderst, Roman. (2008) "Wholesale Price Determination under the Threat of Demand-Side Substitution," University of Frankfurt working paper.

Jeon, Doh-Shin and Jean-Charles Rochet. (2010) "The Pricing of Academic Journals: A Two-Sided Market Perspective," *American Economic Journal: Microeconomics* 2: 222-55.

Lancaster, F. W. and Julie M. Neway. (1982) "The Future of Indexing and Abstracting Services," *Journal of the American Society for Information Science* 33: 83-9.

Lawrence, Steve. (2001) "Free Online Availability Substantially Increases a Paper's Impact," *Nature* 411: 521.

Lubrano, Michel, et al. (2003) "Ranking Economics Departments in Europe: A Statistical Approach," *Journal of the European Economic Association* 1: 1367-1401.

Machado, José A. F. and J. M. C. Santos Silva. (2005) "Quantiles for Counts," *Journal of the American Statistical Association* 100: 1226-37.

McCabe, Mark J and Christopher M. Snyder. (2005) "Open Access and Academic Journal Quality," *American Economic Review Papers and Proceedings* 95: 453-58.

McCabe, Mark J and Christopher M. Snyder. (2007) "Academic Journal Prices in a Digital Age: A Two-Sided Market Model," *The B.E. Journal of Economic Analysis & Policy* 7: Issue 1 (Contributions), Article 2.

McCabe, Mark J and Christopher M. Snyder. (2010) "The Economics of Open Access Journals," Dartmouth College working paper.

Murray, Fiona, Philippe Aghion, Matthias Dewatripont, Julian Kolev, and Scott Stern. (2009) "Of Mice and Academics: Examining the Effect of Openness on Innovation," NBER Working Paper No. 14819.

Palacios-Huerta, Ignacio and Oscar Volij. (2004) "The Measurement of Intellectual Influence," *Econometrica* 72: 963-77.

Papke, Leslie E. and Jeffrey M. Wooldridge. (2008) "Panel Data Methods for Fractional Response Variables with an Application to Test Pass Rates," *Journal of Econometrics* 145: 121-33.

Parker, Kimberly, Kathleen Bauer, and Paula Sullenger. (2003) "E-Journals and Citation Patterns: Is It All Worth It?" *Serials Librarian* 44: 209-13.

Rochet, Jean-Charles and Jean Tirole. (2006) "Two-Sided Markets: A Progress Report," *Rand Journal of Economics* 37: 645-67.

Schonfeld, Roger. (2003) *JSTOR: A History*. Princeton University Press.

Simcoe, Tim. (2008) "XTPQML: Stata Module to Estimate Fixed-Effects Poisson (Quasi-ML) Regression with Robust Standard Errors," Statistical Software Components, Boston College Department of Economics, econpapers.repec.org/RePEc:boc:bocode: s456821.

Tenopir, Carol and Ralf Neufang. (1995) "Electronic Reference Options: Tracking the Changes," *Online* 16: 67-73.

Walker, Thomas. (2004) "Open Access by the Article: An Idea Whose Time Has Come?" *Nature Web Focus* Article 13, April 15.

Wooldridge, Jeffrey M. (1999) "Distribution-Free Estimation of Some Nonlinear Panel Data Models," *Journal of Econometrics* 90: 77-97.

Table 1: Descriptive Statistics

	Level of statistics	Obs.	Mean	Std. dev.	Min.	Max.
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

Notes: Dataset comprised of journal volumes (indexed by v) observed each year (indexed by t) during the citing period. The journal that publishes volume v is denoted $j(v)$.

Table 2: Alternative Specifications for Aggregate Results

Publication years	Citing years	(1)	(2)	(3)	(4)	(5)	(6)
1956-65	1995-99	5.195*** (2.177)	-0.003 (0.071)	-0.007 (0.055)	0.004 (0.033)	-0.061*** (0.013)	0.010 (0.026)
	2000-05	5.589*** (2.409)	0.172 (0.182)	0.123 (0.137)	0.073 (0.051)	-0.007 (0.022)	0.056 (0.044)
		n = 8,944	n = 8,840	n = 8,372	n = 8,372	n = 8,025	n = 8,025
1966-75	1995-99	5.014*** (2.106)	0.033 (0.060)	0.019 (0.059)	-0.010 (0.033)	-0.118*** (0.025)	0.003 (0.027)
	2000-05	3.436*** (1.526)	0.067 (0.094)	0.054 (0.090)	0.025 (0.024)	-0.157*** (0.021)	0.032 (0.029)
		n = 12,506	n = 12,506	n = 12,168	n = 12,168	n = 11,700	n = 11,700
1976-85	1995-99	4.085*** (1.773)	0.041 (0.083)	0.051 (0.086)	0.003 (0.017)	-0.155*** (0.016)	0.031** (0.015)
	2000-05	2.505*** (1.034)	0.051 (0.083)	0.052 (0.084)	0.001 (0.018)	-0.190*** (0.020)	0.025 (0.018)
		n = 18,441	n = 18,441	n = 18,394	n = 18,394	n = 17,618	n = 17,618
1986-95	1995-99	0.855*** (0.324)	0.026 (0.044)	-0.027 (0.021)	-0.010 (0.020)	-0.013 (0.017)	-0.004 (0.017)
	2000-05	1.709*** (0.459)	0.144*** (0.045)	0.109*** (0.032)	-0.008 (0.018)	-0.039** (0.017)	0.007 (0.015)
		n = 15,062	n = 14,907	n = 14,789	n = 14,789	n = 13,815	n = 13,815
1996-2005	1995-99	0.281 (0.225)	-0.073 (0.061)	-0.143*** (0.050)	0.025 (0.048)	0.019 (0.079)	-0.008 (0.045)
	2000-05	1.944*** (0.478)	0.182** (0.077)	0.004 (0.049)	0.026 (0.037)	0.334*** (0.085)	0.009 (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

Notes: Results from Wooldridge's (1999) PQML procedure. 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 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). Bottom of table lists other included variables; "full suite" of time effects refers to the inclusion of publication-year x citation-year fixed time effects. Robust standard errors clustered at the journal level reported in parentheses. Number of observations given at 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 the *10% level, **5% level, ***1% level.

Table 3: Subscription Elasticities for Selected Channels

Publication years	JSTOR: Sole channel (1)	JSTOR: With other channels (2)	Elsevier backfiles (3)
1956-65	0.170**	0.138	†
	(0.072)	(0.089)	
1966-75	0.104**	0.186**	0.023
	(0.047)	(0.078)	(0.065)
1976-85	0.062*	0.068*	-0.049
	(0.036)	(0.036)	(0.054)
1986-95	0.037	0.062**	-0.012
	(0.027)	(0.025)	(0.047)
1996-2005	0.067***	0.017	†
	(0.022)	(0.011)	

Notes: Results from Wooldridge's (1999) PQML procedure. 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 backfiles) 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 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 (access only to part of a volume's content or only for 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-backfile (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 dropped if among ten or fewer for which any online-access indicator is positive. Number of observations approximately that in column 5 of Table 2. Robust standard errors clustered at the journal level reported in parentheses. †No observations fit this category. Significantly different from 0 in a two-tailed test at the *10% level, **5% level, ***1% level.

Table 4: Subscription Elasticities Broken out by Region

	USA			English-speaking West			Non-English speaking West			Rest of world		
	JSTOR			JSTOR			JSTOR			JSTOR		
Publication years	Sole access	With other channels	Elsevier backfiles	Sole access	With other channels	Elsevier backfiles	Sole access	With other channels	Elsevier backfiles	Sole access	With other channels	Elsevier backfiles
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
1956-65	0.183**	0.093	†	0.137	0.223*	†	0.080	-0.182	†	0.286***	0.509***	†
	(0.086)	(0.078)		(0.101)	(0.146)		(0.098)	(0.131)		(0.141)	(0.180)	
1966-75	0.122**	0.175***	0.054	0.080	0.277**	-0.030	0.052	0.087	-0.080	0.145*	0.254**	0.468*
	(0.055)	(0.060)	(0.067)	(0.069)	(0.140)	(0.100)	(0.065)	(0.130)	(0.117)	(0.093)	(0.122)	(0.342)
1976-85	0.056	0.067*	-0.104*	0.042	0.092**	0.053	0.042	0.039	-0.029	0.080*	0.048	0.100
	(0.043)	(0.041)	(0.051)	(0.037)	(0.073)	(0.086)	(0.047)	(0.052)	(0.090)	(0.051)	(0.074)	(0.152)
1986-95	0.039	0.089***	-0.008	0.063*	0.075*	-0.017	-0.014	0.019	-0.052	0.044	0.093**	0.033
	(0.031)	(0.025)	(0.054)	(0.035)	(0.043)	(0.061)	(0.038)	(0.041)	(0.060)	(0.034)	(0.047)	(0.087)
1996-2005	0.081**	0.015	†	0.108*	0.060***	†	-0.249**	-0.075***	†	0.112	0.032	†
	(0.036)	(0.012)		(0.060)	(0.024)		(0.091)	(0.024)		(0.118)	(0.042)	

Notes: Dashed boxes indicate regressions which, to obtain convergence, omit the quadratic term in the journal-age profile and retain only the linear term. Additional specification notes from Table 3 apply here. 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, English-speaking West includes Australia, Canada, Ireland, Israel, New Zealand, and United Kingdom; 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.

Table 5: JSTOR Subscription Elasticities by Quintile

	0-20 Quintile		20-40 Quintile		40-60 Quintile		60-80 Quintile		80-100 Quintile	
Publication years	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)
1956-65	‡		‡		‡		0.323* (0.171)	0.322 (0.320)	0.136* (0.078)	0.105 (0.105)
1966-75	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)
1976-85	-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)
1986-95	-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)

Notes: Quintiles formed by ranking articles within volume by citations in earliest two citing years available (years used for ranking omitted from regressions). Dashed boxes indicate regressions which, to obtain convergence, omit the quadratic term in the journal-age profile and retain only the linear term. ‡Regressions had too few observations with positive citations to produce a non-singular variance-covariance matrix even omitting the linear term in the journal-age profile. See Table 3 for list of additional variables included but not reported, notes about number of observations, specification of standard errors, and definition of symbols. The Elsevier backfile variables reported in Table 3 are also included here but not reported for space considerations.

Table 6: 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 backfiles (3)
1956-65	0.176**	0.200**	†
	(0.069)	(0.081)	
1966-75	0.111***	0.401***	0.071
	(0.039)	(0.056)	(0.085)
1976-85	0.050*	0.078**	-0.020
	(0.027)	(0.036)	(0.045)
1986-95	0.030	0.089***	0.065
	(0.029)	(0.032)	(0.042)
1996-2005	-0.034	-0.024	†
	(0.033)	(0.026)	

Notes: Results from pooled fractional probit estimator developed by Papke and Wooldridge (2008) for panel fractional-response data. Dependent variable is 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 backfiles) interacted with subscribers to those channels, separately interacted with indicators for sole access and full access through some other channel. Coefficients first converted into marginal effects following Papke and Wooldridge's equation (3.10), except we compute the effect at subsample covariate means rather than computing average partial effects. Marginal effects then converted into elasticities by scaling by the ratio of subsample means of the independent variable to that of the dependent variable. Subsample mean of subscribers (rather than the interaction of subscribers with online access) used as scale factor for independent variable. See Table 3 for list of additional variables included but not reported, notes about number of observations, specification of standard errors, and definition of symbols.

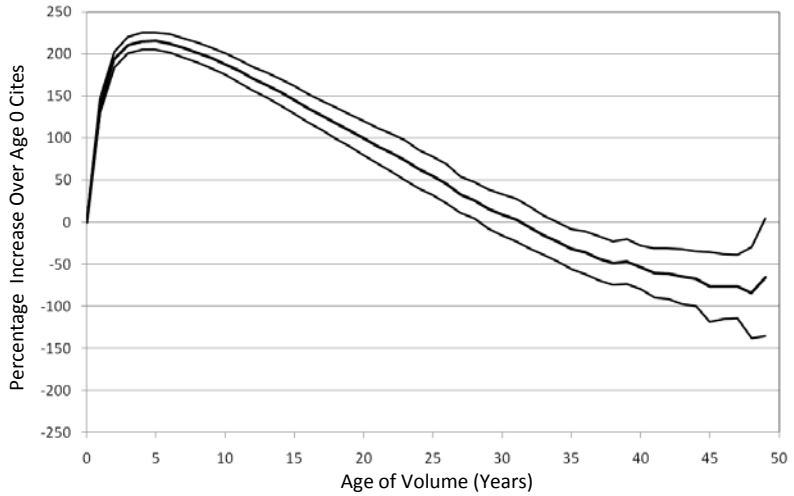


Figure 1: Citation Age Profile. Middle curve plots a set of fixed age effects from Wooldridge's (1999) Poisson quasi-maximum-likelihood procedure. Regression also includes citation-year and journal fixed effects. Outside lines bound 95% confidence interval based on robust standard errors clustered by journal.

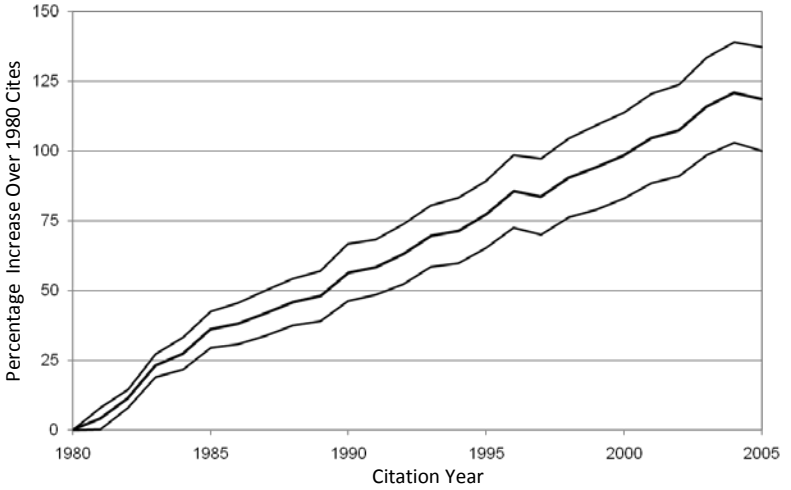


Figure 2: Secular Trend in Citations. Middle curve plots a set of fixed citation-year effects from Wooldridge's (1999) Poisson quasi-maximum-likelihood procedure. Regression also includes a set of age and journal fixed effects. Outside lines bound 95% confidence interval based on robust standard errors clustered by journal.

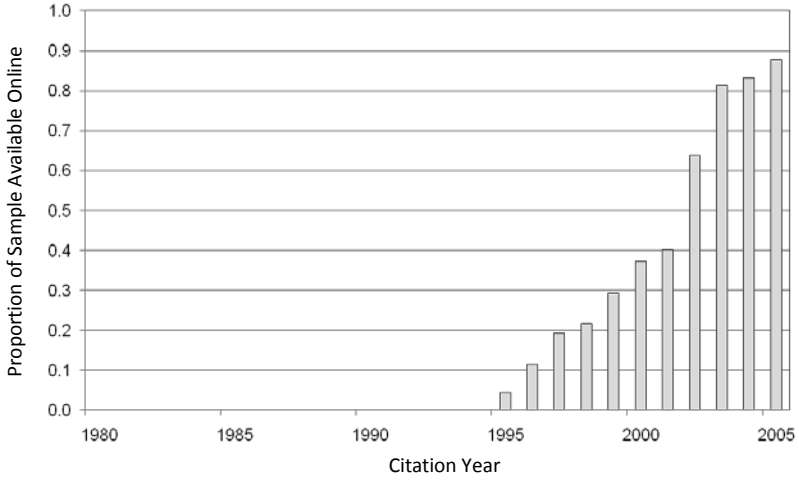


Figure 3: Trends in Online Availability. Mean (across volumes in sample available to be cited in given year) of indicator for online access.

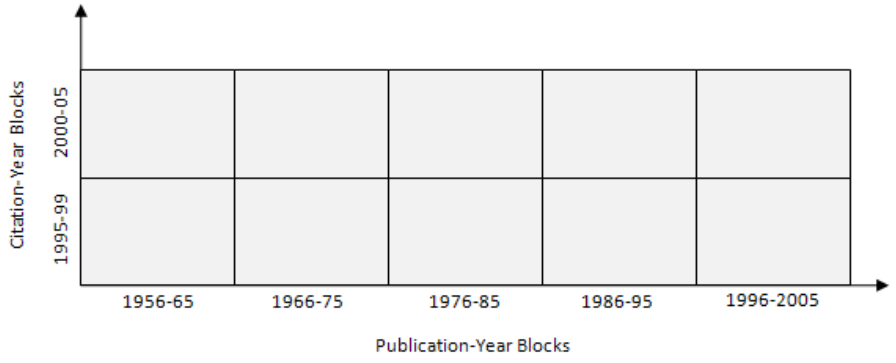


Figure 4: Matrix of Estimated Online Access Effects

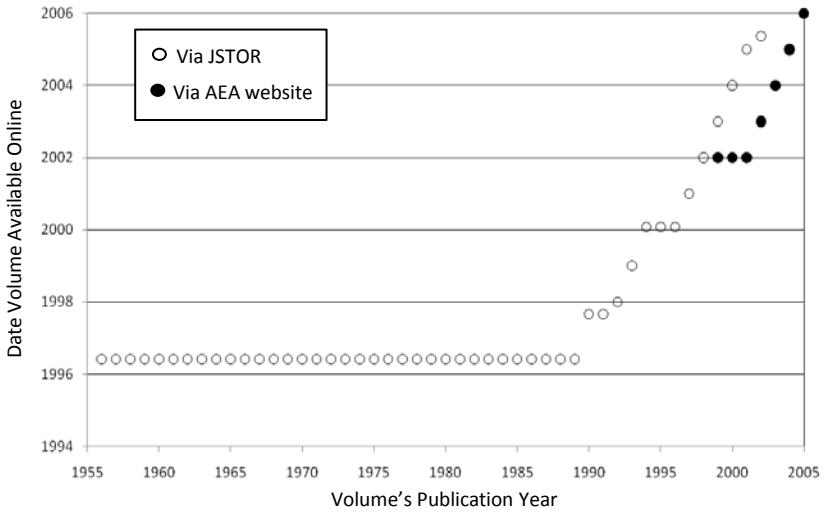


Figure 5: Online Availability in *American Economic Review* Example

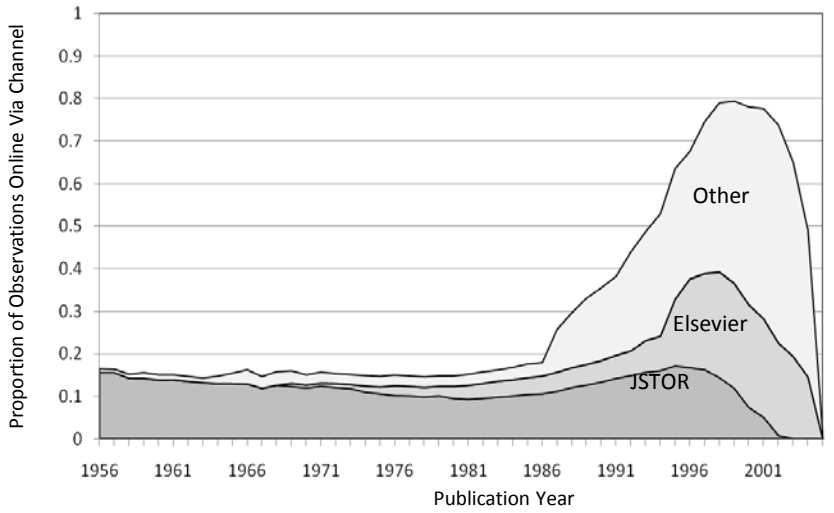


Figure 6: Online Access by Channel for Various Publication Years. Mean value of indicator of online access via JSTOR or Elsevier for each publication year (including both sole access through these channels and duplicate access through some other channel). Means taken across journals and citation years. "Other" is a residual category equal to the mean of an indicator for online access but not through JSTOR and Elsevier.

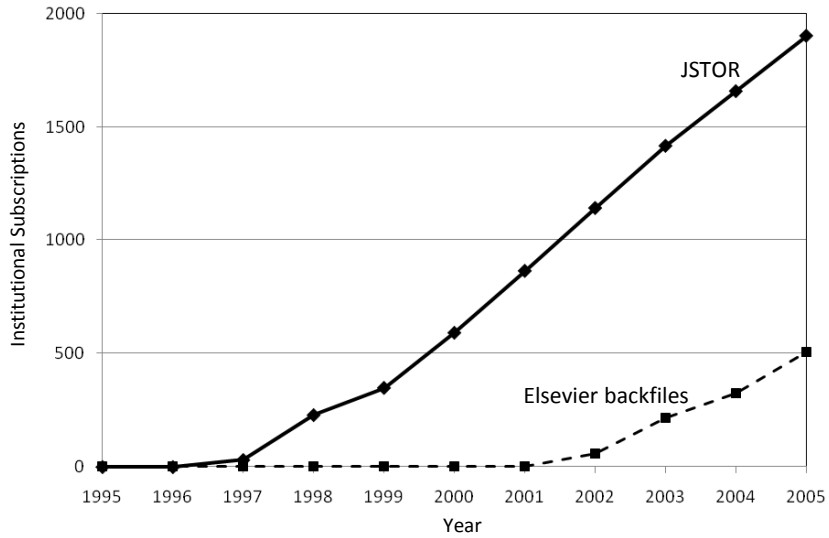


Figure 7: Subscription Trends. Maximum number of institutions subscribing to an online package provided by the indicated channels containing at least one of the journals in our sample. "Elsevier backfiles" refers to institutional purchases of archive of pre-1995 content, which focus on for Elsevier in our subscription analysis.