NBER WORKING PAPER SERIES

CITE UNSEEN: THEORY AND EVIDENCE ON THE EFFECT OF OPEN ACCESS ON CITES TO ACADEMIC ARTICLES ACROSS THE QUALITY SPECTRUM

Mark J. McCabe
Christopher Snyder

Working Paper 28128
http://www.nber.org/papers/w28128

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
November 2020

The authors are grateful to Ted Bergstrom, Laura Braunstein, David Card, Brett Danaher, Barbara DeFelice, Stefano DellaVigna, Robert Johnson, Elizabeth Kirk, Andreas Moxnes, Nina Pavcnik, and participants at the American Economic Association Annual Meetings and International Industrial Organization Conference for helpful comments. The authors thank Mark Bard, Jamie Bergeson-Bradshaw, Yilan Hu, Ella Kim, Scot Parsley, Reagen Readinger, Kyle Thomason, and JasonWei for excellent research assistance. Joseph Brightbill at Clarivate provided the Web of Science category data used in our insider/outsider analysis. Funding for the citation data used in this and earlier papers was supported by a grant from the Andrew W. Mellon Foundation. Work on this paper was supported by a grant from the Alfred P. Sloan Foundation. The authors are grateful for the generous funding from these sources. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Cite Unseen: Theory and Evidence on the Effect of Open Access on Cites to Academic Articles Across the Quality Spectrum

Mark J. McCabe and Christopher Snyder

NBER Working Paper No. 28128
November 2020
JEL No. D83,L17,O33

ABSTRACT

Our previous paper (McCabe and Snyder 2014) contained the provocative result that, despite a positive average effect, open access reduces cites to some articles, in particular those published in lower-tier journals. We propose a model in which open access leads more readers to acquire the full text, yielding more cites from some, but fewer cites from those who would have cited the article based on superficial knowledge but who refrain once they learn that the article is a bad match. We test the theory with data for over 200,000 science articles binned by cites received during a pre-study period. Consistent with the theory, the marginal effect of open access is negative for the least-cited articles, positive for the most cited, and generally monotonic for quality levels in between. Also consistent with the theory is a magnification of these effects for articles placed on PubMed Central, one of the broadest open-access platforms, and the differential pattern of results for cites from insiders versus outsiders to the article's field.

Mark J. McCabe
Questrom School of Business
Boston University
595 Commonwealth Avenue
Boston, MA 48109
and SKEMA Business School, Université Côte d'Azur (GREDEG)
Sophia Antipolis, France
prof.mark.mccabe@gmail.com

Christopher Snyder
Department of Economics
Dartmouth College
301 Rockefeller Hall
Hanover, NH 03755
and NBER
chris.snyder@dartmouth.edu

1. Introduction

Academic journals play a key role in certifying and disseminating research among scholars. An ongoing debate concerns whether this role might be better performed by open-access than traditional journals. Traditional journals earn most of their revenues through library subscription fees, which for commercial publishers have risen to the point that they present a substantial barrier to access.[1] This barrier is removed by open-access journals, which allow free online access to their articles.[2]

Based in part on early studies reporting huge citation benefits from open access[3]—as much as 300%—officials began issuing subsidies and mandates for open-access publication. For example, the European Union issued an open-access mandate for recipients of Horizon 2020 grants, totalling over 80 billion euros over the life of the program. Pending legislation in the United States (the Fair Access to Science and Technology Research Act) mandates open access for grant recipients across a dozen federal agencies.

One of our previous papers, McCabe and Snyder (2014), called into question the validity of these early studies' findings. We showed that such huge estimated effects arise spuriously in a cross section when one fails to account for differences in quality between open- and closed-access content. Using detailed panel data, we generated an estimate of the open-access effect of over 600% in a specification mimicking the early literature by omitting any fixed-effect controls. This

[1] Ted Bergstrom and coauthors have provided a series of studies making the point that prices charged by for-profit journals exceed those charged by non-profit journals by a large factor (Bergstrom 2001, Bergstrom and Bergstrom 2004, Bergstrom and Rubinfeld 2010). For general analyses of journal pricing levels and trends see Dewatripont *et al.* (2006) and Section 2.2.2 of Eger and Scheufen (2018).

[2] For background on various points made in this and the next paragraph, including an overview of the market for academic journals, the open-access debate, the development of open-access journals, and open-access mandates by grant funders, see Chapter 2 of Eger and Scheufen (2018).

[3] Lawrence (2001) found that articles in the proceedings of a computer-science conference available both online and in print received 336% more cites than those available only in print. Harnad and Brody (2004) found that physics articles for which the authors deposited pre-prints on arXiv (an open-access repository of scientific pre-prints) received 298% more cites than those not deposited opn arXiv. Davis and Fromerth (2007) found a 35% citation advantage for mathematics articles deposited on arXiv. Antleman (2004) found that articles for which full-text versions were freely accessible via Google were cited 46% to 91% more than others in across four disciplines. Walker (2004) found that articles in an oceanography journal for which authors purchased hybrid open access received 280% more downloads than others. Eysenbach (2006) studied the effect of open access on citations to articles published in the Proceedings of the National Academy of Sciences, finding as high as a 43% citation boost from open access. See Craig *et al.* (2007) for a survey of the early literature measuring the citation benefit of open access.

estimate was reduced to a modest 8% in our preferred specification controlling for quality with a rich set of fixed effects.

McCabe and Snyder (2014) leveled some of the provocative results of the previous literature but raised some provocative results of its own. The 8% boost from open access was found to be concentrated among the higher-tier journals in the sample; open access led to a significant *reduction* in cites among the lower-tier journals in the sample. That open access could actually reduce cites is surprising and begs explanation. Is it a statistical fluke, perhaps associated with having broken the results down into too many categories or, worse, indicating a problem with the overall methodology? Or is it a systematic outcome, further study of which could contribute to a deeper understanding of the mechanism by which open access boosts citations?

The present paper provides theoretical and empirical support for the latter view. In Section 3, we construct a simple theoretical model offering an explanation of the negative effect of open access for low-quality content. The idea behind the model is that open access facilitates acquisition of full text of the article. The obvious effect is to garner cites from readers who cannot assess its relevance until after reading the full text. There may be a more subtle effect going in the other direction. Some readers may cite articles that they have not read, based on only superficial information about its title or abstract, perhaps rounding out their reference list by borrowing a handful of references gleaned from other sources. If the cost of acquiring the article's full text is reduced by a move to open access, the reader may decide to acquire and read it. After reading it, the reader may find the research a poorer match than initially thought and may decide not to cite it. For the lowest-quality content, the only hope of being cited may be "sight unseen" (pun intended). Facilitating access to such articles may end up reducing their citation counts.

The theory provides a plausible explanation for possible negative effects of open access on citations. Indeed, Monte Carlo exercises based on the theory suggest that negative open-access effects may be the rule rather than the exception. The theory also provides testable predictions that may not have been obvious ex ante but which emerge naturally from the simple model. A distinctive pattern is predicted for the open-access effect across the quality spectrum: the open-access effect should be increasing in quality, ranging from a definitively negative open-access

effect for the worst-quality articles to a definitively positive effect for the best-quality articles. This quality gradient is predicted to be steeper the more convenient is open access. In an extension of the model distinguishing between cites coming from outsiders versus insiders to the article's field, we provide several factors leading outsiders to cite unseen more liberally than insiders. By undoing some of this citing unseen, open access is predicted to reduce outsider citations over a larger range of the quality spectrum than insider citations.

The rest of the paper after Section 3 is devoted to empirical analysis. Sections 4 and 5 describe the data and methods. The sample is the same as in McCabe and Snyder (2014), consisting of citation counts for over 200,000 articles in subfields of science over a ten year period during which open access was an emerging policy. Here we take full advantage of article-level detail, while the previous paper aggregated the data by combining all the articles published by a journal in a year. We rely where possible on the methodology developed in the previous paper, extending it where necessary to accommodate the article-level analysis.

The results are presented in Sections 6 and 7. Section 6 reprises the main provocative results from the previous paper and links them to the theory. Section 7 reports new results breaking articles into quality bins, where the bins are based on cites received by articles during a pre-study period. This provides more detail on how the open-access effect varies across the quality spectrum. As a further test of the theory, we see if a different pattern emerges if an open-access article is also posted on PubMed Central, a huge, open-access repository of science articles offering particularly convenient access. This allows us to test comparative-statics effects with respect to the convenience of open access.

We find that the patterns of the estimates across the quality bins correspond quite closely with those predicted by theory. The open-access effect is roughly monotonic over the quality spectrum. Articles in the lowest-quality bins (receiving zero or one cite in the pre-study period) are harmed by open access; those in the middle experience no significant effect; only those in the top bin with 11 or more cites in the pre-study period experience a benefit from open access. Moving from open access through the journal's own website to open access through PubMed Central pivots the open-access effect so that it is even more sensitive to quality, resulting in greater losses to low-quality

articles and greater gains to high-quality articles. PubMed Central access reduces cites to articles in the zero- or one-cite bins by around 14% while increasing cites to articles in the bin with 11 or more cites by 11%.

When we divide cites by their source, from insiders versus outsiders to the article's field, we find different patterns over the quality spectrum. Inside cites show the same pattern of negative open-access effects in low-cite bins and positive in high-cite bins as we find in our basic results, but the positive effects are observed for a broader range of bins, even those with moderate numbers of cites. By contrast, the effect of open access on outsider cites is negative across virtually the entire quality spectrum.

2. Literature Review

Our paper is part of the literature that attempts to identify the causal effect of open access on citations more carefully by moving beyond simple cross-sectional regressions.[4] Davis *et al.* (2008) conduct a field experiment, randomly selecting articles from American Physiological Society journals for open access, finding little causal effect. Gaule and Maystre (2011) take an instrumental-variables approach, instrumenting for authors' endogenous decision to pay a $1,000 fee to have their *Proceedings of the National Academy of Sciences* (*PNAS*) articles openly accessible using the timing of budget cycles, again finding little causal effect. Evans and Reimer (2009) and McCabe and Snyder (2014) take a panel-data approach to identification of the causal effect. Incorporating fixed effects, they measure the change in cites to a given journal volume over time as it moves from closed to open access compared to control volumes that do not change during that period. Both find about an 8% average effect of open-access. Mueller-Langer and Watt (2018) look at the effect of hybrid access (open access at a gated journal purchased by the author) using a similar

[4] In addition to the early studies reviewed in footnote 3, more recent cross-sectional studies include Atchison and Bull (2015) in political science and Tang, Bever, and Yu (2017) in ecology. Piwowar, *et al.* (2018) study a cross section of 100,000 articles across many disciplines, attempting to control for quality by including article age and field variables. Li, *et al.* (2018) (see also related work by Yan and Li 2018) study a panel of journals over time, allowing them to estimate the effect of a change in access status for a given journal compared to control journals that do not change access status over the period. This identification strategy may not hold quality constant because the contents of the journal changes over time. The move to open access may attract a different caliber of article; the move may signal broader changes to journal operations including altering editorial standards.

difference-in-differences approach, but in their case the difference in open access stems from a policy granting authors at certain institutions free hybrid access at Springer journals. They find no effect of hybrid access for articles having freely accessible pre-prints but a 6–8% citation boost for other articles. Staudt (2020) takes a difference-in-differences approach comparing National Institutes of Health (NIH) funded articles to matched controls before and after an NIH issued an open-access mandate, estimating around a 4% open-access effect. Bryan and Ozcan (2020) study the same mandate but look at citations in subsequent patents rather than academic articles, finding a 27% effect. The papers cited in this paragraph employ convincing empirical strategies for causal identification, but none provides theory or empirical results across articles of different qualities as does the present paper.

The theoretical section of our paper contributes to the growing theoretical literature on the economics of open access, including Shavell (2010); Mueller-Langer and Watt (2010); Mueller-Langer and Scheufen (2013); McCabe, Snyder, and Fagin (2013); McCabe and Snyder (2015); and McCabe and Snyder (2018). Most closely related are models that focus on heterogeneity in quality across articles, including McCabe and Snyder (2005), McCabe and Snyder (2007), Jeon and Rochet (2010), Armstrong (2015), Scheufen (2015), Feess and Scheufen (2016), and Besancenot and Vranceanu (2017). A novel aspect of the present model is that we separate the option to cite from the option to read, allowing for the emergence of a new strategy: citing unseen.

The present paper is related to the broader literature looking for "long tail" or "superstar" effects of Internet distribution. Recent studies suggest that online retailing boosts sales more for products in the long tail, in markets ranging from clothing (Brynjolfsson, Hu, and Simester 2011) to video sales (Elberse and Oberholzer-Gee 2008). McCabe and Snyder (2015) find that the increase in citations from moving from print to digital access through JSTOR is fairly uniform across article qualities. We share an interest in measuring potential heterogeneity in effects between popular and unpopular items, but we study a change in the price of Internet access rather than the move from off-Internet to on-Internet distribution. The theoretical model and findings that open access disproportionately benefits superstar content and reduces cites in the long-tail are new in the present paper.

3. Theory

3.1. Model

At the outset of the game, we take as given the existence of a continuum of citable articles. Let $q \in [0,1]$ denote a given article's quality, a random variable that has a distribution over the population of articles characterized by density function $\phi(q)$. Let $\bar{q} = \int_0^1 q\,\phi(q)\,dq$ denote the expected value of q. We model quality as the probability that the article is relevant to a reader's research. More formally, relevance characterizes the match between the given article and an individual reader, a Bernoulli random variable equaling 1 (success) with probability q and 0 (failure) with probability $1-q$. An article may be low quality because it is badly written, poorly executed, on a narrow topic, or some combination of these reasons.

Articles attract cites from a continuum of readers with mass normalized to 1. Readers have different types corresponding to the gross benefit $b \geq 0$ each receives from citing relevant research. Assume b is continuously distributed on its support, $(0,\infty)$. Let P be the probability measure associated with b. While we do not deny the possibility that a large mass of readers have $b = 0$ in practice, such valueless readers end up ignoring the article in equilibrium in the model, so are irrelevant to the analysis.

Readers are the only strategic players in the game. Readers can apply one of three strategies to each article. First, they can simply ignore the article. Second, they can cite the article without reading it, based on its title, abstract, or other superficial information—the strategy referred to as "citing unseen" in our title. Third, they can acquire the article's full text, read it, and issue a deep cite if it turns out to be relevant. Our baseline model assumes readers choose strategies before learning the article's q, knowing only the distribution of q over the population of citable articles.

Normalize a reader's payoff from not citing an article to $V_n = 0$. His or her expected payoff from citing unseen is

$$V_u = \int_0^1 [qb-(1-q)s]\phi(q)\,dq = \bar{q}b-(1-\bar{q})s. \tag{1}$$

An article of quality q provides an expected gross benefit of bq, the value to the reader of type b of citing relevant research times the probability q an article of that quality is relevant. Subtracted off

of this expected gross benefit is the expected cost of leaving a bad impression on one's audience when one cites poor or irrelevant research. This expected cost equals the probability $1-q$ that the reader happens to have cited an irrelevant article times the sanction $s > 0$ that his or her audience issues in that event.[5]

A reader's expected payoff from acquiring the full text of the article is

$$V_f = \int_0^1 [q(1+\theta)b]\phi(q)\,dq - a = \bar{q}(1+\theta)b - a. \tag{2}$$

Equation (2) differs from (1) in three ways. First, the cost $a > 0$ of acquiring the full test of the article has been subtracted. This cost includes any associated fees, the hassle and delay for articles that are not immediately available on the reader's computer screen after a few mouse clicks, and the effort involved in reading and digesting the article. Second, readers who acquire the full text of the article learn whether the article is a match for their research. This allows them to avoid the sanction s from citing an irrelevant article. Third, deep cites have more value for readers' research than superficial cites, captured by the factor $\theta > 0$ multiplying the gross benefit.

3.2. Equilibrium Citations

Let $B_n(a)$ be the set of reader types who strictly prefer not to cite a given article in equilibrium, $B_u(a)$ the set of types who strictly prefer to cite unseen, and $B_f(a)$ the set of types who strictly prefer to acquire the full text. Recalling the normalization $V_n = 0$, we have

$$\begin{aligned} B_n(a) &= \{b \,|\, 0 > \max\{V_u, V_f\}\} \\ B_u(a) &= \{b \,|\, V_u > \max\{0, V_f\}\} \\ B_f(a) &= \{b \,|\, V_f > \max\{0, V_u\}\}. \end{aligned} \tag{3}$$

The arguments of $B_n(a)$, $B_u(a)$, and $B_f(a)$ emphasize their dependence on a, stemming from the dependence of V_f on a shown in equation (2). The specification of all inequalities as strict in (3)

[5] It is straightforward to add a physical cost c of composing the citing passage and bibliography entry. We omit c from the presentation because the results are qualitatively similar and because, in practice, the physical cost is likely to be much smaller than the reputational harm from citing irrelevant research.

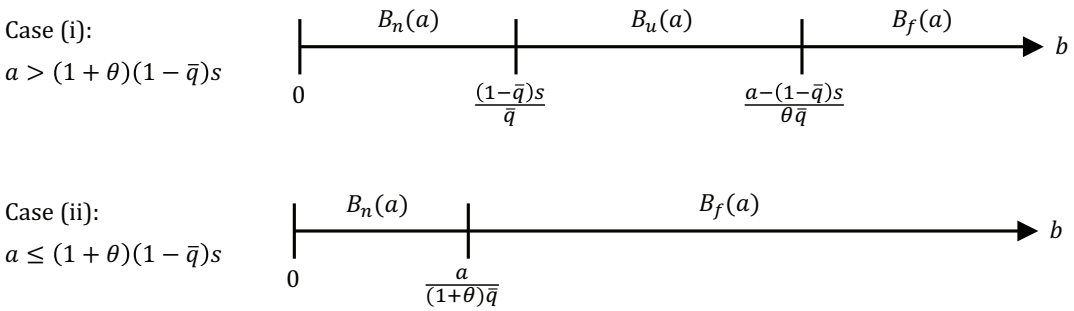

Figure 1: Sets of Reader Types Pursuing Various Citing Strategies. $B_n(a)$ denotes the set of reader types b preferring not to cite, $B_u(a)$ the set preferring to cite unseen, and $B_f(a)$ the set preferring to acquire full text. See equation (3) for formulas.

is without loss of generality because the distribution of types is continuous, implying indifferent types have zero measure.

Combining equations (1)–(3), straightforward algebra can be used to show that $B_n(a)$ is a non-empty interval containing the lowest values of b and that $B_f(a)$ is a non-empty interval containing the highest values of b. Thus, $B_u(a)$ must form an interval between $B_n(a)$ and $B_f(a)$. It is straightforward to show that interval $B_u(a)$ has positive measure if and only if $a > (1+\theta)(1-\bar{q})s$.

Figure 1 provides the precise formulas for the bounds between sets $B_n(a)$, $B_u(a)$, and $B_f(a)$. The figure presents the two cases depending on whether $B_u(a)$ has positive measure. To gain some intuition for the equilibrium, focus on the richer of the two cases, case (i). We see that an interval of the lowest types do not cite the article; their gross benefit is too low to justify any sanction or acquisition costs. Higher types in the next set engage in the strategy of citing unseen. Their gross benefit from citing is high enough to risk the sanction of citing an article that turns out to be irrelevant. The highest types acquire the full text because their marginal benefit, θb, from citing deeply rather than superficially is high enough to justify the high acquisition cost that defines case (i). In case (ii), the low acquisition cost that defines this case leads any reader who decides to cite the article to acquire the full text and cite it deeply. Hence, no reader engages in citing unseen in this case.

Having characterized readers' equilibrium citing behavior, it is a simple matter to tally up the cites received by a given article. The article receives no cites for each type in $B_n(a)$ and one cite for each type in $B_u(a)$. For each type in $B_f(a)$, the article is cited with the probability q that the

reader finds the article to be relevant after seeing the full text. Combining these considerations, the mass $x(q,a)$ of cites received by an article of quality q is

$$x(q,a) = P(B_u(a)) + qP(B_f(a)). \tag{4}$$

A key feature of equilibrium is that a reader who acquires the full text generates fewer expected cites than one who cites unseen. This will generate some of the counterintuitive comparative-static effects investigated next.

3.3. Open-Access Effect

In this subsection, we study the comparative-static effect on an article's citation count, $x(q,a)$, of a move from closed to open access. We model the move from closed to open access as a decrease in acquisition cost from a_c to $a_o < a_c$. Open access decreases acquisition cost for readers at institutions who did not subscribe to the fee-based journal: those readers would have to pay a charge (typically \$40 as of this writing) for the article or go through the interlibrary loan process and wait for the article to be delivered. Even for readers at subscribing institutions, the move to open access facilitates access the article, relieving the reader from the series of steps needed to access the article via library resources.

The move from closed to open access can have a variety of effects depending on which case from Figure 1 is relevant. To focus on interesting outcomes, we assume that a_c is high enough to tempt some readers to cite unseen. More formally, the following parametric assumption is maintained throughout the analysis:

$$a_c > (1+\theta)(1-\bar{q})s. \tag{5}$$

It is straightforward to prove that (5) is necessary and sufficient for a positive measure of readers to engage in citing unseen under some regime in the model.[6]

[6]Assume (5) holds. Then (i) is the relevant case from Figure 1 under closed access. In this case, $B_u(a_c)$ is a non-degenerate subinterval of $(0,\infty)$ and so has positive measure since the distribution of types is assumed to have support $(0,\infty)$. Hence, a positive measure of types engage in citing unseen under some regime. Assume instead that

Proceeding with the analysis, condition (5) ensures that, under closed access, (i) is the relevant case from Figure 1. The move from closed to open access may leave the model in case (i) or may shift the case to (ii), leaving us with two possibilities to analyze depending on the value of a_o. Assume first that that $a_o > (1+\theta)(1-\bar{q})s$, so that (i) is the relevant case under both closed and open access. Then, the reduction in acquisition cost from a_c to a_o leaves the boundary between $B_n(a_c)$ and $B_u(a_c)$ the same but shifts the boundary between $B_u(a_c)$ and $B_f(a_c)$ to the left. Some types who cited unseen under closed access switch to acquiring the full text under open access. This switch reduces the article's citation count because citing unseen generates one cite per reader but acquiring the full text only generates a cite with probability q. For the highest quality articles ($q = 1$), the conversion of types from citing unseen to acquiring the full text does not affect total citations because all readers who acquire the full text are certain to find it relevant and cite it. For all other articles of quality $q < 1$, however, the move from closed to open access strictly reduces cites. The reduction is greater the lower is the value of q because the types who have been converted into acquiring the full text are less likely to result in cites when q is low.

Next, assume $a_o \leq (1+\theta)(1-\bar{q})s$. The move from closed to open access changes the relevant case from (i) to (ii), generating two effects. All the types that were in $B_u(a_c)$ switch into $B_f(a_o)$. This switch contributes to a reduction in the article's cites since acquiring the full text is less likely to generate a cite than citing unseen. Another effect is that $B_f(a_o)$ extends beyond the boundary of $B_u(a_c)$, cutting into $B_n(a_c)$. This implies that some of the types that did not cite under closed access acquire the full text under open access, contributing to an increase in the article's cites. The two effects work in opposite directions; which dominates depends on the quality q of the cited article. If q is low, the article loses more cites from types switching from citing unseen to acquiring the full text since the full text is unlikely to generate a relevant cite for low q. The opposite is true if q is high. Hence, when $a_o \leq (1+\theta)(1-\bar{q})s$, the effect on citations is monotonically increasing in q, negative for q near 0, and positive for q near 1.

The preceding analysis provides an intuitive sketch of the following proposition, proved for-

(5) is violated. Then (ii) is the relevant case under closed access. It is also the relevant case under open access since $a_o < a_c$. Thus, both $B_u(a_c)$ and $B_u(a_o)$ have zero measure, implying that no positive measure of readers engages in citing unseen.

mally in the appendix. To make the statement of the proposition precise, some further notation is required. Let $\Delta(q, a_c, a_o)$ denote the marginal effect on cites when an article of quality q moves from closed to open access, measured as a proportional change:

$$\Delta(q, a_c, a_o) = \frac{x(q, a_o) - x(q, a_c)}{x(q, a_c)}. \tag{6}$$

Subscripts on $\Delta(q, a_c, a_o)$ denote partial derivatives with respect to those arguments.

Proposition 1. *Assume* (5) *holds. For all* $q \in [0,1]$, $\Delta_q(q, a_c, a_o) > 0$ *and* $\Delta_{qq}(q, a_c, a_o) < 0$. *If* $a_o \geq (1+\theta)(1-\bar{q})s$, *then* $\Delta(q, a_c, a_o) < 0$ *for all* $q \in [0,1)$ *and* $\Delta(1, a_c, a_o) = 0$. *If* $a_o < (1+\theta)(1-\bar{q})s$, *then* $\Delta(q, a_c, a_o) < 0$ *for* q *sufficiently close to* 0 *and* $\Delta(q, a_c, a_o) > 0$ *for* q *sufficiently close to* 1.

The proposition states that the open-access effect $\Delta(q, a_c, a_o)$ is an increasing, concave function of article quality q. Though the sign of the open-access effect is ambiguous, the proposition makes a strong case that negative values are not anomalous. For all parameters, the open-access effect is negative for some articles—a range of the lowest-quality ones. For some parameters, the open-access effect is negative for almost all articles (i.e., except for a set of zero measure).

Figure 2 illustrates the proposition with bin-scatter plots from a Monte Carlo exercise. Details behind the exercise including the distributions and parameters used are provided in the figure notes. The dots are averages of the open-access effect $\Delta(q, a_c, a_o)$, plotted at the midpoint of each of twenty equal-sized quality bins. The only difference between the two curves regards the parameter a_o, with the black curve setting a high value for a_o and the gray curve setting a low value for a_o.

All the claims made in the proposition are illustrated in the figure. Both curves are strictly increasing and concave. The reader can verify that the parameters behind the black curve satisfy $a_o > (1+\theta)(1-\bar{q})s$. Consistent with the proposition in this case, the black curve lies below the horizontal axis for all $q < 1$, approaching the horizontal axis as q approaches 1. The reader can verify that the parameters behind the grey curve satisfy $a_o < (1+\theta)(1-\bar{q})s$. Consistent with the proposition in this case, the grey curve crosses the horizontal axis, becoming positive for sufficiently high values of q. Notice that, for the quite standard functional forms and parameters that we have chosen for the Monte Carlos, negative values of the open-access effect are the rule rather than the exception.

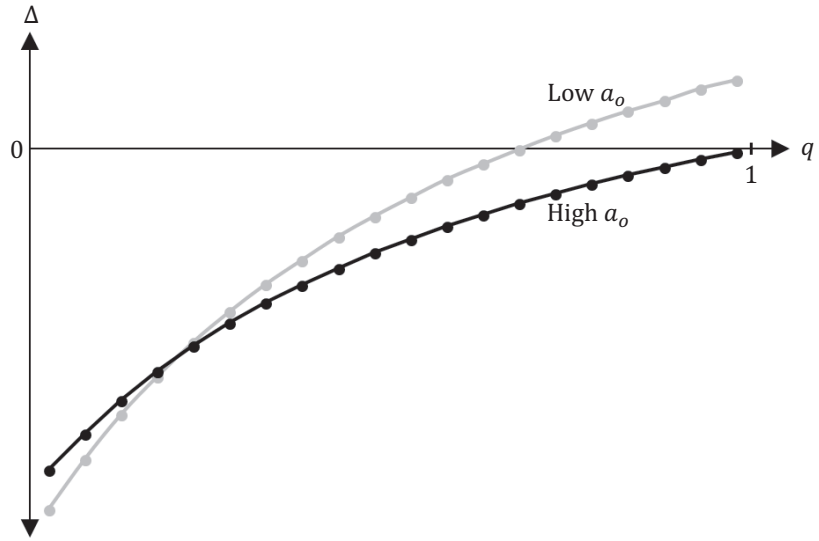

Figure 2: Monte Carlo Exercise Illustrating Basic Results. Dots are averages of the open-access effect $\Delta(q, a_c, a_o)$, plotted at the midpoint of each of twenty equal-sized quality bins constructed from 10 million Monte Carlo draws. The exercise uses a uniform $[0,1]$ distribution for q and a standard half normal for b, implemented by taking the absolute value of a random draw from a normal distribution with mean 0 and standard deviation 1. The exercise sets $\theta = 2.5$, $s = 0.25$, and $a_c = 1$, parameters which the reader can verify satisfy (5). The only difference between the two curves regards the value of a_o, with $a_o = 0.5$ for the black curve and $a_o = 0.1$ for the gray curve. The same 10 million draws were used for each curve.

3.4. Platform Convenience

This subsection explores how the open-access effect varies with the convenience of the open-access platform. While the pecuniary costs associated with open access are the same on any platform—access by definition being free—non-pecuniary costs may differ across open-access platforms. For example, free access via PubMed Central may be more convenient than via the journal's website because it is a huge archive that readers may be accustomed to using, and thus involves more efficient access. We allowed for possibly positive values of a_o to reflect such non-pecuniary costs. To reflect the differential non-pecuniary costs of different open-access platforms, we will introduce two possible values of the open-access acquisition cost: the cost associated with a broad open-access platform, a_o^B, and with a narrow one, a_o^N, where $a_o^B < a_o^N$. We use the labels "broad" and "narrow" to provide a concrete idea of why one open-access platform may involve lower non-pecuniary costs than another, but we have in mind any factors that make open-access platform b more convenient than platform n.

The next proposition states the intuitive result that when the move from closed to open access

entails a larger reduction in acquisition costs, the open-access effect is intensified. Geometrically, the graph of $\Delta(q, a_c, a_o)$ as a function of q rotates counterclockwise, becoming even more steeply positively sloped. The magnitude of the open-access effect experienced at both extremes of quality is exaggerated, becoming weakly more negative for the lowest-quality ($q = 0$) articles and weakly more positive for the highest-quality ($q = 1$) articles. The appendix contains a formal proof.

Proposition 2. *Assume* (5) *holds. Consider an increase in the convenience of the open-access platform, lowering a_o^N to $a_o^B < a_o^N$. The slope of the open-access effect strictly increases: $\Delta_q(q, a_c, a_o^B) > \Delta_q(q, a_c, a_o^N)$. The negative open-access effect for the lowest-quality articles becomes weakly more negative: $\Delta(0, a_c, a_o^B) \leq \Delta(0, a_c, a_o^N) < 0$. The non-negative open-access effect for the highest-quality articles becomes weakly more positive: $\Delta(1, a_c, a_o^B) \geq \Delta(1, a_c, a_o^N) \geq 0$.*

Figure 2, drawn to illustrate various cases in Proposition 1, serves to illustrate Proposition 2 as well. The curves are identical except for the assumed value of a_o. The move from the black to the grey curve can be thought of as an increase in the convenience of open access. The black curve shows the open-access effect from a narrow platform with access cost a_o^N and the grey curve from a broad platform with lower access cost, $a_o^B < a_o^N$. As predicted by the proposition, the move from the black to the grey curve is a counterclockwise pivot, exaggerating the negative effect for q near 0 and exaggerating the positive effect for q near 1.

3.5. Insiders Versus Outsiders

This subsection models differences in the citing behavior of readers depending on whether they are insiders—coming from the same discipline as the article they may cite—or outsiders—coming from a different discipline. These differences in citing behavior may then translate into differences in the open-access effect. We will model two main differences.

It is reasonable to suppose that the sanction s for an irrelevant cite depends mostly on the perceptions of knowledgeable scholars in the discipline. These scholars must know both the cited and the citing articles, typically requiring the articles to come from the same discipline, in turn requiring that the citing reader be an insider in the cited article's discipline. A reader who is an outsider is less likely to suffer much of a sanction. Scholars with the knowledge to judge irrelevance may not interact enough with the citing reader to have their opinion matter. Scholars

in the same discipline as the citing reader would be in a poor position to judge the relevance of his or her cite to an unfamiliar discipline. We model this difference by assuming that the sanction is higher for an insider than an outsider. More specifically, to simplify the analysis while at the same time approximating reality, we suppose that the positive sanction $s > 0$ is fixed for insiders; for outsiders we take the limit as the sanction vanishes: $s \to 0$.

The analysis for insiders is identical to what we found before since we have fixed the same positive sanction for them. All the various cases from Proposition 1 continue to be possible.

To analyze the case of outsiders, observe that after imposing the limit $s \to 0$ in Proposition 1, the outcome is pinned down to case (i), and moreover $B_n(a)$ becomes vanishingly small. The only effect of a reduction in a for outsiders is to shift some readers from citing unseen to acquiring the full text, which must reduce cites. Before stating this result formally in the next proposition, we introduce some further notation. Let superscripts O and I indicate variables related to outsiders and insiders, respectively. Thus $B_n^O(a)$, $B_u^O(a)$, and $B_f^O(a)$ denote the sets of outsider types who do not access the article, cite it unseen, or acquire the full text. Let $x^O(a,q) = P^O(B_u^O(a)) + qP^O(B_f^O(a))$ denote the number of cites from outsiders, where P^O is the conditional probability measure on outsider types b. Let $\Delta^O(q,a_c,a_o) = [x^O(q,a_o) - x^O(q,a_c)]/x^O(q,a_c)$ denote the open-access effect on outsider cites. Define P^I, $x^I(q,a)$, and $\Delta^I(q,a_c,a_o)$ analogously for insiders. The following proposition is an immediate corollary of Proposition 1.

Proposition 3. *The open-access effect for outsiders, never positive, is negative for all but the highest-quality article:* $\Delta^O(q,a_c,a_o) < 0$ *for all* $q \in [0,1)$ *and* $\Delta^O(1,a_c,a_o) = 0$.

Note that the right-hand side of condition (5) equals 0 when $s = 0$, implying that (5) holds automatically for outsiders; so there is no need to explicitly assume the condition in the statement of Proposition 3.

Thus far, insiders are identical to the generic readers of the baseline model. A possible difference is that well-informed insiders may have considerable information about an article before accessing it, perhaps based on a general familiarity with articles in their discipline, perhaps based on an ability to make sharp inferences *ex tempore* from an article's author, title, or abstract information within one's discipline. We model this feature formally by assuming that a proportion

$\sigma \in (0,1]$ of insiders are "smart," able to observe q before accessing the article; the remaining $1-\sigma$ proportion are ordinary insiders who as before only know the distribution of q but not the realized value for the article before accessing it.

Analysis of smart insiders is more complicated than before because the number of cites is no longer linear in q as it was in equation (4). The sets $B_u^I(a)$ and $B_f^I(a)$ become functions of q for readers who can see q before deciding on their strategy. Some general results for extreme values of q are still available, stated in the next proposition.

Proposition 4. *Assume* (5) *holds. For the lowest-quality articles, the open-access effect for insiders is the same negative value as for generic readers:* $\Delta^I(0,a_c,a_o) = \Delta(0,a_c,a_o) < 0$. *For the highest-quality articles, the open-access effect for insiders is lower than that for generic readers (i.e.,* $0 < \Delta^I(1,a_c,a_o) < \Delta(1,a_c,a_o)$*) unless* $\Delta(1,a_c,a_o) = 0$ *in which case* $0 = \Delta^I(1,a_c,a_o) = \Delta(1,a_c,a_o)$.

The proof in the appendix again works closely with Figure 1. The difference for smart insiders is that the known value q must be substituted everywhere in the figure for the mean $\bar{q}$.

Figure 3 provides bin-scatter plots for a new Monte Carlo exercise illustrating possible outcomes of the insider/outsider model. Details behind the exercise are provided in the figure notes. The dashed curve shows the open-access effect for ordinary insiders, equivalent to generic readers in this variant of the model. They experience a negative effect for low q and positive effect for high q. The grey curve represents the open-access effect for outsiders. They are formally identical to ordinary insiders except that the sanction has been reduced from $s = 0.3$ to $s = 0$ for them. As expected from Propostion 3, the black curve is everywhere below the horizontal axis and approaches the axis as q approaches 1.

The black curve represents the open-access effect for insiders when some of them are smart. The sanction for these readers has been returned to the original positive level for regular insiders but now a fraction $\sigma = 0.75$ of them are smart and can see the value of q for articles before acquiring the full text. As expected from Proposition 4, the curve approaches that for the regular insiders as q approaches 0 and is between the generic reader's curve and the horizontal axis as q approaches 1. For values of q between 0 and 1, the figure illustrates the possibility of a highly non-monotonic open-access effect, in this example rising above the horizontal axis for a "pocket"

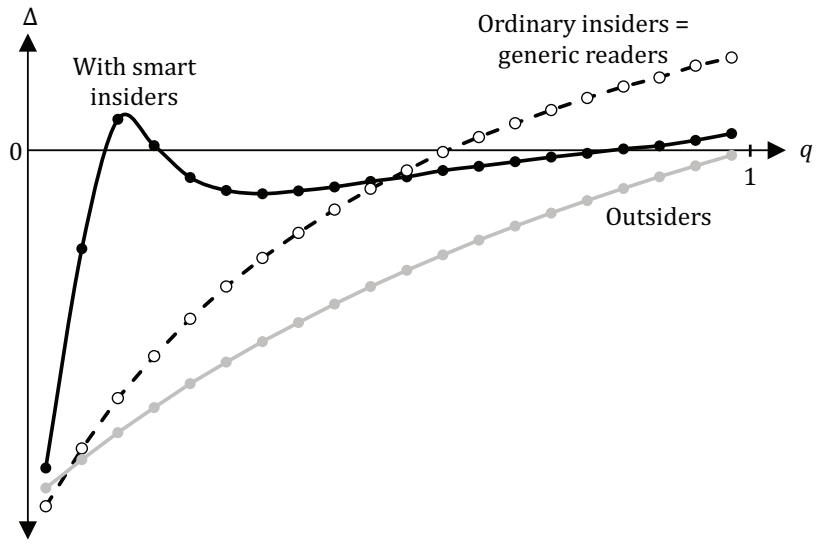

Figure 3: Monte Carlo Exercises Illustrating Insider/Outsider Analysis. As in previous figure, dots are averages of the open-access effect $\Delta(q, a_c, a_o)$, plotted at the midpoint of each of twenty equal-sized quality bins constructed from 10 million Monte Carlo draws. Here, rather than the baseline model used in the previous figure, reader behavior is governed by the insider/outsider model. The distributions and parameters are the same as in the previous figure except for two differences. To emphasize certain features of the curves, s has been adjusted slightly, from $s = 0.25$ to $s = 0.3$. The fraction of smart insiders is set to $\sigma = 0.75$.

of some relatively low values of q, dipping back below for larger q, and rising above the horizontal axis again for the highest values of q. While such a drain-pipe shape is not guaranteed—indeed, the curve reverts to the monotonic gray curve in the limit $\sigma \to 0$—Figure 3 documents the possibility. The theoretical possibility that the open-access effect is highest for a "pocket" of moderate rather than the highest quality articles hinges on the presence of smart insiders in the model. Such insiders are smart enough to avoid citing mediocre articles unseen. For them, the main effect of a move from closed to open access is to increase the measure obtaining full access, which can translate into a large open-access effect Δ since Δ is measured as a percentage increase over a potentially very small base of cites that a mediocre article would receive from smart insiders under closed access.

Overall, our insider/outsider analysis has several empirical implications. We expect the open-access effect for outsiders to be negative across the quality spectrum. The effect should also be negative for insiders citing the lowest-quality articles. Insiders may exhibit positive open-access effects for higher quality articles; the effect may exhibit non-monotonicities and may be highest for articles of intermediate rather than the highest quality.

3.6. Other Extensions

In the model, facilitating access to the article's full text can reduce cites by giving the reader a more precise signal of the article's quality, possibly reducing cites to low-quality articles. This is one possible mechanism behind a negative open-access effect, but other mechanisms might be possible having similar empirical implications.

The move from closed to open access was modeled as reducing the reader's access cost from a_c to a_o. The move could entail another, indirect effect, explored in a first extension. The acquisition cost falls not only for the reader but also for the reader's audience. Thinking of s as the *expected* sanction received conditional on citing an irrelevant article—multiplying the reputational harm conditional on the irrelevant cite being discovered times the probability of discovery—when the cost of acquiring the cited article falls for the reader's audience, this increases the probability they discover the irrelevant cite. A move from closed to open access in this extended model increases the expected sanction from s_c to $s_o > s_c$ in addition to reducing access cost, $a_o < a_c$, for the reader. The next proposition implies that the higher sanction depresses cites to open-access articles. Let $x(q, a_c, s_c)$ and $x(q, a_o, s_o)$ denote cites, respectively, to a closed- and open-access article in this extended model, where the added argument emphasizes dependence on the expected sanction, which can vary under closed and open access.

Proposition 5. *Let $s_o'' > s_o'$ in the extended model allowing sanctions to be greater under open access. Then $x(q, a_o, s_o'') \leq x(q, a_o, s_o')$, with strict inequality if and only if*

$$a_o > (1+\theta)(1-\bar{q})s_o'. \tag{7}$$

Let $\Delta(q, a_c, s_c, a_o, s_o) = [x(q, a_o, s_o) - x(q, a_c, s_c)]/x(q, a_c, s_c)$ denote the open-access effect in the extended model. Since $x(q, a_c, s_c)$ is independent of s_o, it is immediate from Proposition 5 that $\Delta(q, a_c, s_c, a_o, s_o)$ is nonincreasing in s_o and strictly decreasing if and only if (7) holds. Thus, a negative open-access effect only becomes more likely if open access has the added indirect effect of increasing sanctions.

Another way open access could reduce cites is by intensifying competition among articles to be cited. For example, acquiring the full text of an article may increase the chance that readers

encounter a substitute article in the reference list that they prefer to cite as the basis for a particular idea. Alternatively, the platforms or search techniques used by readers when looking for open-access articles may lead them to notice more related articles than when looking for closed-access articles. Either way, lower quality articles are more likely to lose in the competition. If open access intensifies an article's exposure to competition, it can reduce cites, especially to low-quality content.

The insider-outsider model assumed that the key difference between outsiders and ordinary insiders in that outsiders face little sanction for citing irrelevant content. This effect can be amplified if it is assumed in addition that outsiders obtain a lower average value $\bar{q}$ from citing because the probability of a good match is lower. Refer to Figure 1, in particular case (i), the relevant case for outsiders when $s = 0$. A decrease in $\bar{q}$ shifts the boundary between $B_u(a)$ and $B_f(a)$ right. For very low values of $\bar{q}$, virtually all outsider citing under closed access would be unseen, leading the move to open access to generate an even larger reduction in cites.

The model has the unrealistic implication that certain types of reader may end up citing all articles. To avoid this implication, the population of articles could be reinterpreted as being just those in a narrow topic area rather than all academic articles. Alternatively, we could add an awareness function, $A(q)$, equal to the probability that the reader knows about the existence of the article, a necessary condition to be cited. Then no reader would end up citing all articles, just the smaller set of those of which he or she is aware. The awareness function would factor out of Δ and not affect any of our implications for the open-access effect.

4. Data

Our analysis is based on the sample of 100 science journals used in McCabe and Snyder (2014). The sample is built around the subfield of ecology, which we selected among the subfields of hard science for several reasons: (a) it is a well-defined subfield, (b) it involves a manageable number of journals, and (c) it experienced substantial growth in open access. The sample includes all of the ecology journals in Thomson ISI's set of indexed journals. This accounts for 60% of the journals in the sample. Of the remaining 40%, 60% were taken from botany, the most closely related subfield

to ecology, and 40% from multidisciplinary science and biology, presuming that some ecology and botany research is published in such general-interest journals. We selected the top journals from these latter two categories, ranked based on the standardized ISI yearly impact factors averaged over the period 1985-2004. We restricted the overall number of journals to 100 because of the considerable expense and effort involved for each additional journal. The appendix provides a table listing the sample journals by field.

The dataset merges citations data together with historical information on online and open accessibility. The citations data was acquired from Thomson ISI. For each of the 100 journals in our sample, ISI lists every article published since 1996. Each published article is linked to all cites from the over 8,000 ISI-indexed journals for each year from 1996 to 2005. To this basic citation data we merged hand-collected information on whether the full-text article was available online or open access. To determine online availability, we determined the date on which each journal issue was placed online either on the journal's own website or one of the major digital aggregators by contacting publishers and aggregators, cross-checking their reports using libraries' electronic journal catalogs and the Internet Archive (www.archive.org). We collected information on open access for each issue in a similar way.

The resulting dataset from these two sources includes observations for over 230,000 individual cited articles, indexed by i. Let $j(i)$ index the journal, $v(i)$ the volume, and $p(i)$ the year in which article i is published. Our dataset has a panel structure because each article receives cites each year, from the year it is published until the end of the sample period in 2005. Let t index the citation year. Note the distinction between the dataset's two time indexes: $p(i)$ indexes the year the *cited* article was published, while t indexes the year the *citing* article was published. Because the average article comes out in the middle of the ten-year sample period and hence has five years of observations for citations, the full panel contains over 1.2 million article-citation-year observations. The articles in our sample received 4.8 million cites from ISI-indexed articles over the sample period.

Table 1 provides descriptive statistics for the dataset. One average, articles received 3.78 cites per year. The standard deviation of yearly cites (11.69) is quite high, as is the range, from a low of 0 (the case for nearly half of the observations) to a high of 1,145 (received the year after

Table 1: Descriptive Statistics for Combined Sample

Variable	Mean	Median	Std. Dev.	Min.	Max.
Publication year	1998.99	1999.0	2.47	1996	2005
Citation year	2002.00	2002.0	2.46	1996	2005
Age	3.00	3.0	2.46	0	9
Cites in year	3.78	1.0	11.69	0	1,145
Top-ranked journal indicator	0.72	1.0	0.45	0	1
Online-access indicators					
• Partial	0.18	0.0	0.38	0	1
• Full	0.68	1.0	0.47	0	1
Open-access indicators					
• Partial	0.09	0.0	0.29	0	1
• Full	0.18	0.0	0.38	0	1
PubMed-access indicators					
• Partial	0.04	0.0	0.20	0	1
• Full	0.07	0.0	0.26	0	1

Notes: Panel dataset consists of 1,268,386 observations for 231,407 articles.

publication by the 2002 article in *Nature* describing the sequencing of the human genome). The next row reports statistics for an indicator for top-ranked journals, constructed by ranking journals within our sample using the same impact factor used to select journals for inclusion in the sample, and then setting the indicator to 1 if the article appears in a journal in the top half of this ranking. While this indicator divides the number of journals in half by design, many more than half of the article-level observations (72% to be precise) appear in top-ranked journals because these tend to publish more articles each year.

The last four rows of Table 1 provide information on indicators for online and open access. For 68% of the observations, article i was available online through some channel for the full year t. For a much smaller fraction of observations, 18%, article i was openly accessible for the full year t. We will focus on full online and full open access throughout the analysis. The regressions will also include indicators for partial online and partial open access—set to 1 if some of the content in the volume containing the article was available in the indicated way for some of the year but not all of the content for the full year—but we will not focus on those results because partial access is a catch-all category combining observations with various durations of access.

5. Methodology

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:

$$E\left(Cites_{it} \,\middle|\, Age_{it}, Access_{it}, p(i), j(i)\right) = \exp\left(\alpha_i + \beta_{p(i)t} + \gamma^1_{j(i)} Age_{it} + \gamma^2_{j(i)} Age^2_{it} + \delta Access_{it}\right). \quad (8)$$

$Cites_{it}$ denotes the number of cites to article i in year t, $Age_{it} = t - p(i)$ denotes the article's age, and $Access_{it}$ denotes a vector of variables capturing the nature of access to the article. The remaining variables are parameters to be estimated: α_i is an article fixed effect, $\beta_{p(i)t}$ is a time effect possibly varying for each publication year $\times$ citation year combination, $\gamma^1_{j(i)}$ and $\gamma^2_{j(i)}$ are coefficients on a quadratic age profile separately estimated for each journal, and δ is a vector of parameters capturing access effects. Wooldridge (1999) provides a Poisson quasi-maximum-likelihood (PQML) estimator for equation (8), which, as long as the conditional mean is specified correctly, produces consistent estimates of the parameters for any positive conditional distribution of $Cites_{it}$ (Poisson, negative binomial, or other).

Including article fixed effects, α_i, in (8) helps remove the bias that plagued cross-sectional studies of the open-access effect cited in the introduction. If higher quality articles are more likely to be openly accessible, the open-access coefficient in previous studies may just be picking up quality differences between open- and paid-access articles. The time effects $\beta_{p(i)t}$ help control for several patterns widely observed in citation data—documented in the present dataset by McCabe and Snyder (2015)—that could otherwise confound the results. McCabe and Snyder (2015) document the hump-shaped path of citations for an article as it ages, peaking two to three years after publication (in the case of the sampled science articles) before gradually petering out. It is important to control for the age profile to avoid, for example, confounding the natural decline in citations after age three with the effect of open access that might have started then. McCabe and Snyder (2015) also document a rising secular trend in citations, likely reflecting the growth over time in the number of indexed journals, articles per journal, and cites per article. It is important to control for this secular growth to avoid confounding it with the effects of online and open access, both of which become

more prevalent later in the sample. The included time effects $\beta_{p(i)t}$ are flexible enough to control for a distinct citation age profile for each yearly cohort of articles having an arbitrary shape. The $\beta_{p(i)t}$ are simultaneously flexible enough to control for secular growth in citations that differs by vintage of content and that has an arbitrary shape. The quadratic age profile involving coefficients $\gamma^1_{j(i)}, \gamma^2_{j(i)}$ contributes to the flexibility by controlling for a citation age profile with a possible hump shape that differs across journals within a cohort. McCabe and Snyder (2015) demonstrated the importance of including this rich set of controls, showing that omitting one or another of them produced unreliable results.

Indicators for online and open access, both partial and full, are contained in the $Access_{it}$ vector in equation (8). Two challenges must be overcome for our access indicators to provide consistent estimates. First, the access variables must be exogenous, i.e., orthogonal to the error given by the difference between observed citations and the conditional expectation in (8). Given the wealth of controls included in the conditional expectation, the nature of access to an article is plausibly orthogonal to the remaining error. Consider the example of *Plant Physiology*, shown in Figure 4. In 2001, the journal allowed open access to a whole tranche of volumes through 1999. After that, the journal maintained a policy of making articles available open access after a two-year embargo behind a pay wall. This pattern of maintaining a fixed embargo period combined with episodes in which a tranche of back issues is made openly accessible is fairly typical and seems to be based more on technological convenience than on departures from the expected pattern of cites to a volume.

The second challenge is that the access variables must exhibit some independent variation from the other regressors. For example, if each volume of a journal were made openly accessible after the same embargo duration after publication, the open-access indicator would be completely collinear with the volume's age. As Figure 4 shows, this is not typically the case. Paradoxically, the tranche of 1996–99 volumes that *Plant Physiology* made openly available in 2001 helps identify the effect of open access on cites because simultaneously turning on the open-access indicator hits different volumes at different points in their age profiles. The 1996 volume is first openly accessible in its fifth year after publication, the 1997 in its fourth year, the 1998 in its third, and

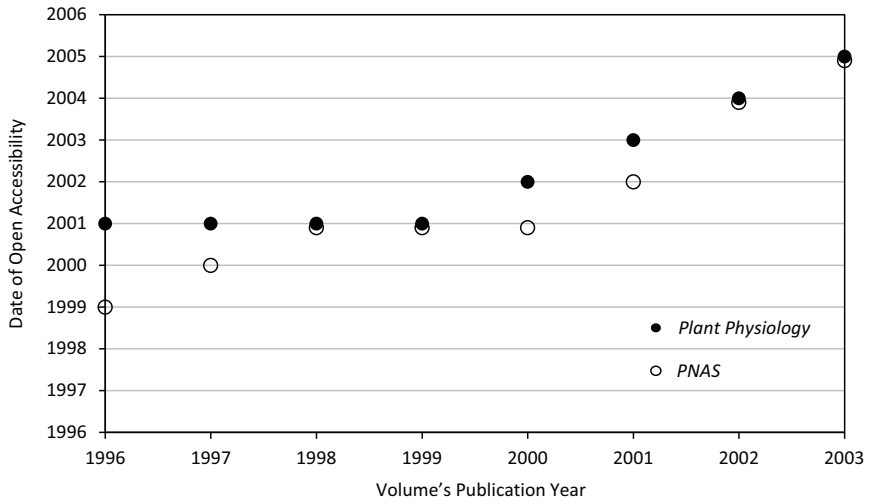

Figure 4: Patterns of Open Accessibility for Example Journals. Shows the earliest full year of open access for volumes of two journals in our sample, *Plant Physiology* and *PNAS*.

so forth. The 1996 volume provides information on what the citation age profile should look like through the fourth year in the absence of open access. If the 1998 volume deviates from this pattern in 2001, say experiencing a jump relative to expectations, this jump can be attributed to the effect of the start of open access in that year. For this identification strategy to be valid, one must be able to purge secular time effects using data from other journal volumes of the same vintage but having a different pattern of open access. Our data satisfy this requirement. First, most journals in our sample are never openly accessible, providing a natural control sample. For journals that have some open access in our sample, the timing of open access follows idiosyncratic patterns. In the case of *PNAS*, the 1996 and 1997 volumes were already open access by 2001. *PNAS* granted open access to slightly different tranche of volumes in 2001 than *Plant Physiology*.

6. Aggregate Results

In this section we reprise some of the provocative results from McCabe and Snyder (2014) that motivated the theory presented in Section 2. Table 2 presents the coefficients of interest from specifications of a count-data model along the lines of equation (8), allowing various interactions with the full-open-access indicator in different columns. Although the regressions contain all the controls listed in the table notes, to conserve space we only report the results for open-access in-

Table 2: Open-Access Results for Combined Sample

Variable	(1)	(2)	(3)
Partial open access	0.044*	0.046*	0.039*
	(0.024)	(0.024)	(0.023)
Full open access	0.081***		
	(0.027)		
(a) Interacted with top-50 journal		0.085***	
		(0.027)	
(b) Interacted with bottom-50 journal		–0.185***	
		(0.059)	
(c) Interacted with no PubMed access			0.072***
			(0.085)
(d) Interacted with PubMed access			0.046
			(0.033)
Test of interactions conducted		(a) = (b)	(c) = (d)
χ^2 test statistic		17.1***	7.3**
Articles	154,744	154,744	154,744
Panel observations	941,795	941,795	941,795
Article fixed effects	Yes	Yes	Yes
Publication × citation year fixed effects	Yes	Yes	Yes
Partial online-access indicator	Yes	Yes	Yes
Full online-access indicator	Yes	Yes	Yes
Journal-specific age profile	Quadratic	Quadratic	Quadratic

Notes: Each column is a different specification of a regression using Wooldridge's (1999) PQML procedure. Dependent variable is cites to an article in a citation year. Results converted into marginal effects $\exp(\beta)-1$, where β is the Poisson regression coefficient and $\exp(\beta)$ is the incidence rate ratio. Robust standard errors clustered at the journal level reported in parentheses. Reported sample size is smaller than in Table 1 since we report observations remaining after dropping articles contributing no identifying variation to coefficients of interest in presence of article fixed effects, both articles having only one year of citation data and articles receiving no cites over the sample period (in which case dependent variable is constant within the fixed-effect group). Significantly different from 0 in a two-tailed test at the *10% level, **5% percent level, ***1% level.

dicators. The reported standard errors are robust to heteroskedasticity and clustered at the journal level. Regression coefficients have been converted into marginal effects interpretable as proportionate increases.

The results in column (1) show that full open access increased cites by 8.1%, statistically significantly different from 0 at the 1% level. The effect of partial open access is about half that of full access, 4.4%, consistent with partial access averaging to about half a year of access.[7] This is

[7]Although the regressions in Table 2 are run on article-level data with article fixed effects while the comparable

the set of results leading us to conclude that open access causes a small positive boost to citations on average in the sample.

The remaining columns break the results down into categories to look for sources of heterogeneity. Column (2) allows the marginal effect of full open access to differ between the 50 top-ranked journals in our sample and the remaining 50. The marginal effect for the top-50 journals, 8.5%, is similar to the basic result we obtained before dividing journals by rank. The marginal effect for the bottom-50 journals provides the surprise motivating the present paper: it is significantly negative, with open access leading to a 18.5% reduction in cites for these journals. An explanation consistent with the theory is that articles in the bottom-50 journals tend to be to the left of the quality spectrum in Figure 2, experiencing a negative open-access effect, while articles in the top-50 tend to be to the right of the quality spectrum, experiencing a positive open-access effect.

In column (3) we estimated separate marginal effects for open access solely through the journal's own website on the one hand and through PubMed Central (in addition to the journal's website) on the other. While access solely through a journal website continues to have a significantly positive effect, additional access through a potentially broader platform (PubMed) is significantly smaller and indeed is not significantly different from zero. This is surprising at first glance because it is natural to suppose if open access boosts cites on average, open access through a yet more convenient channel should boost cites even more. Here, we are seeing the opposite: more convenient access appears to reverse the benefits from open access. An explanation consistent with theory is that offering PubMed access in addition to open access through the journal's website is akin to moving from a narrow to a broad open-access platform, reducing non-pecuniary costs of accessing the article. This results in a pivot in the curve capturing the open-access effect in Figure 2 from the light to the dark curve. Both losses and gains are exaggerated, but if losses are exaggerated more (technically, if the probability-weighted integral between the curves to the left of their intersection is greater than to the right), then the overall effect of adding PubMed access could be to reduce the

regressions in McCabe and Snyder (2014) are run on considerably more aggregate volume-level data with volume fixed effects, the two sets of results are identical. The results are guaranteed to be identical because the other regressors besides the fixed effects are the same in both sets of regressions and these other regressors do not vary across articles within a volume.

average estimated open-access effect. The regression in column (3) is too crude to test this claim, so we turn to more detailed results next.

7. Results for Quality Bins

7.1. Baseline Results

This section provides more detailed estimates of the effect of open access for articles at different points in the quality spectrum. A traditional approach would be to apply quintile regression, minimizing the sum of asymmetrically weighted absolute residuals to yield estimates of specific quintiles. Unfortunately, while this method has been extended to the case of count data (Machado and Silva 2005), no such estimator has been developed for panel count data.

Our version of quintile analysis consists of estimating equation (8) separately for each of five quality bins formed on the basis of their citation counts. In order to avoid bias due to selection of the sample based on residuals, we use different data when constructing these quality bins than we use to estimate the regressions. We use citations in the first two years after publication (called the "selection period") to form the quality bins but run the regression using cites in the third and later years (called the "regression period"). We later report results where the bins are based on percentiles (quintiles) of cites, but our preferred procedure will form five bins based on absolute numbers of cites. This is our preferred procedure for forming the bins because enough articles have the same number of cites, especially at the low end of zero, one, or two, that quintile bins end up dividing articles with the same number of cites period into different quintiles at random. Fortunately, we will see that the results are robust to binning procedure.[8]

Five quality bins were formed: articles with no cites in the first two years after publication, articles with one cite, articles with 2–5 cites, articles with 6–10 cites, and the remaining articles with 11 or more cites. Table 3 provides descriptive statistics on cites in the ex post period in each

[8]McCabe and Snyder (2015) discuss conditions under which binning based on citations in an ex ante period does not lead to bias. Although based on a different sample—business and economics rather than science journals—extensive analysis there showed that the results were similar whether one, two, or three years of cites were included in the ex ante period and whether gaps of various lengths were allowed between ex ante and ex post periods. We use the same length for the ex ante period (two years) and for the gap (zero years) as the preferred procedure in that paper.

Table 3: Descriptive Statistics for Citation Bins

Cites in selection period	Obs.	Cites each year over regression period				
		Mean	Median	Std. Dev.	Min.	Max.
0 cites	361,657	0.53	0.0	1.28	0	59
1 cite	126,426	1.66	1.0	2.27	0	50
2-5 cites	183,734	3.54	3.0	3.77	0	87
6-10 cites	67,745	7.63	6.0	6.35	0	126
11+ cites	89,106	24.96	17.0	31.90	0	1,071

of these bins. Not surprisingly, cites are strongly correlated across the two periods, as shown by the increase in mean cites reading down that column.

Table 4 reports the results from estimating equation (8) separately for each of these five bins. Each column of the table represents a separate regression. As in Table 2, the regressions include a rich set of controls, but for space considerations we only report coefficients for the open-access variables of interest. The controls are identical to those used in Table 2 with one exception. Table 2 included quadratic age profiles for each journal to help control for hump-shaped citation age profiles specific to each journal. The sample used for Table 4 excludes the first two years of citation data used to select the bins. Given that the peak of the citation age profile is reached two years after publication, the profile remaining after truncating the first two years is roughly linear, adequately captured by linear age profiles for each journal. We found the time series remaining after dropping the first two years of data for each article to be too short to reliably estimate quadratic age profiles.

The results for partial open access are again about half of those for full open access fairly consistently throughout the table, so the discussion will focus on the full-open-access results. The result is negative for the first three bins and positive for the last two. The –12.3% marginal effect in the 1-cite bin and the 7.7% marginal effect in the 11+ cite bin are both significant at the 1% level. Figure 5 provides a clearer picture of how the results compare to theory. Each dot is the plot of the estimated full-open-access effect from Table 4 on the vertical axis against median cites for each bin during the selection period reported in Table 3 on the horizontal axis. The curve is a quadratic fit to the points, providing a summary view of the pattern of the results. The fitted curve displays many

Table 4: Open-Access Results by Citation Bins

	Cites in selection period				
Variable	0 cites	1 cite	2–5 cites	6–10 cites	11+ cites
Partial open access	0.000	–0.083***	–0.017	0.001	0.034***
	(0.041)	(0.030)	(0.034)	(0.021)	(0.010)
Full open access	–0.046	–0.123***	–0.045	0.010	0.077***
	(0.060)	(0.026)	(0.040)	(0.025)	(0.013)
Articles	31,008	21,271	35,245	13,496	17,775
Panel observations	162,735	107,622	173,469	65,357	86,371
Article fixed effects	Yes	Yes	Yes	Yes	Yes
Publication × citation year fixed effects	Yes	Yes	Yes	Yes	Yes
Partial online-access indicator	Yes	Yes	Yes	Yes	Yes
Full online-access indicator	Yes	Yes	Yes	Yes	Yes
Journal-specific age profile	Linear	Linear	Linear	Linear	Linear

Notes: Each column is a separate regression including observations falling into indicated bin. Bins formed by summing cites in selection period (first two years after publication). Observations in selection period are omitted from the regressions, reducing the sample size relative to that reported in Table 1. Sample sizes are further reduced by dropping articles contributing no identifying variation to coefficients of interest in presence of article fixed effects, both articles having only one ex post year of citation data and articles receiving no cites over the ex post period (in which case dependent variable is constant within the fixed-effect group). Regressions use Wooldridge's (1999) PQML procedure. Dependent variable is cites to an article in a citing year. Results converted into marginal effects $\exp(\beta)-1$, where β is the Poisson regression coefficient and $\exp(\beta)$ is the incidence rate ratio. Robust standard errors clustered at the journal level reported in parentheses. Significantly different from 0 in a two-tailed test at the $^{*}10\%$ level, $^{**}5\%$ percent level, $^{***}1\%$ level.

of the features predicted by theory: it starts out below the horizontal axis for the lowest quality articles and eventually rises above it for the highest quality articles. It is monotonically increasing and concave.

Tables B2 and B3 provided in Appendix B demonstrate the robustness of the results to alternative structures for the bins. Like Table 4, Table B2 forms bins based on absolute numbers of citations but uses different numbers to form the bins: 0 cites, 1–2 cites, 3–9 cites, and 10+ cites. This scheme results in one fewer bin. The results are still consistent with Table 4. The effect of full open access is negative for the lowest bins and positive and significant only for the highest bin. Rather than binning based on absolute number of cites during the selection period, Table B3 bins articles based on cites relative to other articles published in the same year. The bins are asymmetric. The first bin is the largest, containing articles in the 0-50 percentile. All the articles had

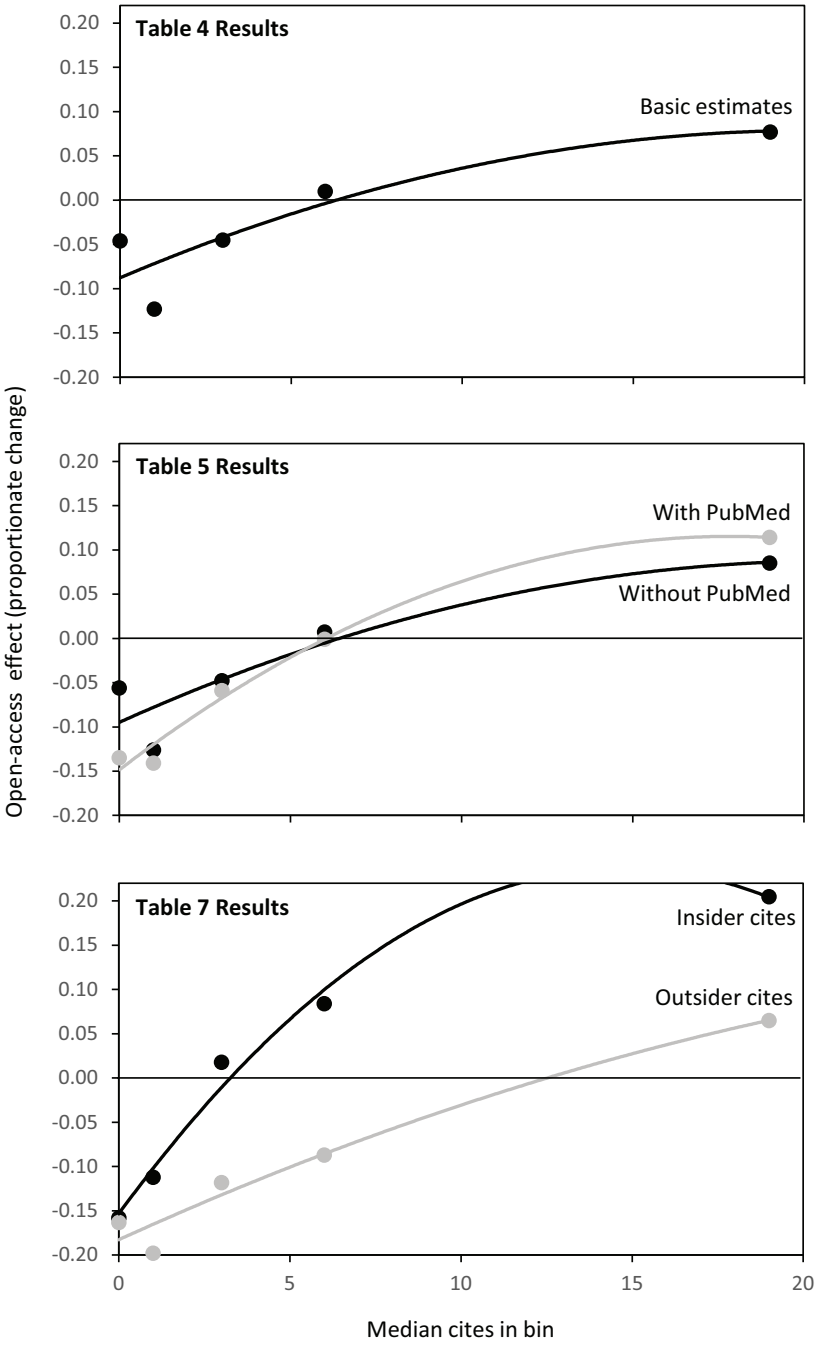

Figure 5: Plot of Results by Citation Bins. Plots results from separate regressions by citation bin collected from several tables. Each dot is the plot of the open-access effect for a citation bin against against median cites in that bin.

no cites in the selection period in this percentile, so there was no reason to divide it further. The remaining four percentiles have equal 12.5% widths for symmetry. The results in Table B3 are nearly identical to those in each of the corresponding bins in Table 4. We conclude that the results are robust to different binning procedures.

7.2. Interactions with PubMed Access

Table 5 separates the effect of full open access into access solely through the journal's own website in row (a) and additional access through PubMed in row (b). Access solely through the journal's own website produces almost identical estimates to those for full open access that are not separated by channel in Table 4. The results that are interacted with PubMed access are noticably different. The negative effects in the three lowest-quality bins are exaggerated as is the positive effect for the highest-quality bin. Now articles in the 0-cite bin suffer a 13.5% decline in cites from this enhanced form of open access, significant at the 10% level. The articles in the 11+ cite bin gain 11.4%, significant at the 1% level. The open-access results interacted with PubMed access differ significantly from those interacted with no PubMed access at the 1% level for the lowest and highest cite bins.

The pattern of the results is again best seen in a graph. The results from Table 5 are plotted in the middle panel of Figure 5. The results interacted with no PubMed access are plotted as the black dots and fit by the black quadratic curve; the results interacted with PubMed access are plotted as the grey dots and fit by the grey quadratic curve. As predicted by theory—see, in particular, Figure 2—the increase in convenience of access when the article is posted on PubMed pivots the curve so that the lowest-quality articles are harmed even more by this form of access and the highest-quality articles benefit even more.

Given that the coefficients line up fairly well with the fitted quadratic curves through them and that the quadratic curves resemble those used to illustrate the theory, we have a fairly strong demonstration that the empirical results support the model's theoretical predictions. The one place where the results depart from the theory is in the non-monotonicity in the effect of full open access moving from the 0-cite to the 1-cite bin. In the initial results abstracting from PubMed access

Table 5: Results by Citation Bins Interacted with PubMed Access

Variable	0 cites	1 cite	2–5 cites	6–10 cites	11+ cites
Partial open access	–0.007	–0.085***	–0.019	0.000	0.035***
	(0.045)	(0.029)	(0.034)	(0.020)	(0.009)
Full open access					
(a) Interacted with no PubMed access	–0.056	–0.126***	–0.048	0.007	0.085***
	(0.064)	(0.027)	(0.041)	(0.023)	(0.008)
(b) Interacted with PubMed access	–0.135*	–0.141***	–0.059	–0.001	0.114***
	(0.066)	(0.036)	(0.047)	(0.022)	(0.007)
χ^2 statistic from test of (a) = (b)	18.7***	1.1	0.7	1.2	7.2***
Articles	31,008	21,271	35,245	31,215	17,775
Panel observations	162,735	107,622	173,469	156,875	86,371
Article fixed effects	Yes	Yes	Yes	Yes	Yes
Publication × citation year fixed effects	Yes	Yes	Yes	Yes	Yes
Partial online-access indicator	Yes	Yes	Yes	Yes	Yes
Full online-access indicator	Yes	Yes	Yes	Yes	Yes
Journal-specific age profile	Linear	Linear	Linear	Linear	Linear

Notes: See Table 4 for applicable notes.

reported in Table 4, the marginal effect dips from -4.6% to –12.3% before rising again to –4.5% in the 2–5 cite bin. The results interacted with no PubMed access exhibit similar non-monotonicity. The non-monotonicity is less pronounced in the results on the interaction with PubMed access. There, there is hardly a dip between the –13.5% effect in the 0-cite bin to the –14.1% effect in the 1-cite bin.

The disaggregated results by quality bin in Table 4 dovetail with the aggregate results from Table 2. This pattern that open access reduces cites in lower bins and increases cites in higher bins echoes the findings from column (2) of Table 2 that open access reduced cites for the bottom-50 journals and increased cites for the top-50 journals. The PubMed effect estimated in column (3) of Table 2 for the combined sample—small and not significantly different from 0—is consistent with negative open-access effects observed for low-quality articles (for which the grey curve in the middle panel of Figure 5 falls below the horizontal axis) averaging out positive effects observed for high-quality articles (for which the grey curve rises above the horizontal axis).

Table 6: Descriptive Statistics for Insider/Outsider Analysis

Variable	Mean	Median	Std. Dev.	Min.	Max.
Cites in year in total	1.80	1.0	3.41	0	133
Cites in year from insiders	1.00	0.0	2.17	0	83
Cites in year from outsiders	0.76	0.0	1.66	0	80

Notes: Descriptive statistics for subsample of botany and ecology journals in our dataset, which are amenable to measuring insider and outsider cites. Resulting panel has 622,500 observations for 114,961 articles.

7.3. Insider Versus Outsider Cites

This subsection presents open-access estimates broken down by whether cites are coming from insiders versus outsiders relative to the cited journal's field. The first step in identifying insiders and outsiders is to restrict our cited sample by eliminating multidisciplinary-science journals. These 17 journals cover such broad areas that nearly all scientists could be regarded as insiders to them. The remaining 83 titles are in more narrow fields of ecology and botany. For a botany journal, a cite is classified as inside botany if and only if the citing journal has botany as one of its ISI-designated subjects. An analogous procedure is used to classify cites as inside or outside ecology.

Table 6 provides descriptive statistics for the restricted sample used in the insider/outsider analysis. Excluding multidisciplinary-science titles cuts the number of articles and panel observations about in half compared to Table 1. The loss of observations is disproportionate to the number of titles cut, which is expected since the typical multidisciplinary-science journal publishes considerably more articles per year than a more specialized botany or ecology journal. The average botany or ecology journal obtains 1.80 cites per year in total across inside, outside, and ambiguous sources, fewer than the 3.78 cites in our original sample including multidisciplinary-science journals, reflecting the fact that the field journals in our sample obtain fewer cites per article than the more general, multidisciplinary-science journals. On average, the restricted sample receives 1.00 insider cites (56% of total) and 0.76 outsider cites (42% of total). The residual 2% of cites come from multidisciplinary science and popular science titles, which cannot be unambiguously classified as outsider or insider.

Table 7 reports estimates of the marginal effect of open access by citation bins where the depen-

Table 7: Insider/Outsider Results by Citation Bins

Variable	Cites in selection period				
	0 cites	1 cite	2–5 cites	6–10 cites	11+ cites
	A. Insider cites				
Partial open access	–0.144**	–0.096***	0.015	0.080	0.149**
	(0.055)	(0.033)	(0.025)	(0.094)	(0.081)
Full open access	–0.158***	–0.112***	0.018	0.084**	0.205***
	(0.031)	(0.038)	(0.023)	(0.040)	(0.083)
Articles	24,014	13,420	14,377	2,226	678
Panel observations	125,261	67,678	69,967	10,479	3,059
	B. Outsider cites				
Partial open access	–0.065**	–0.108*	–0.040	–0.080	–0.223***
	(0.031)	(0.054)	(0.054)	(0.069)	(0.073)
Full open access	–0.163***	–0.198***	–0.118***	–0.087	0.065
	(0.044)	(0.040)	(0.029)	(0.076)	(0.083)
Articles	30,095	13,499	10,678	913	156
Panel observations	157,040	67,603	51,256	4,185	729
Article fixed effects	Yes	Yes	Yes	Yes	Yes
Publication × citation year fixed effects	Yes	Yes	Yes	Yes	Yes
Partial online-access indicator	Yes	Yes	Yes	Yes	Yes
Full online-access indicator	Yes	Yes	Yes	Yes	Yes
Journal-specific age profile	Linear	Linear	Linear	Linear	Linear

Notes: Each column of each of panels A and B reports a separate regression including observations falling into bin indicated in column heading. Only insider cites are used for the dependent variable in panel A and only outsider cites in panel B. The same subset of cites used for the dependent variable is used to form bins: i.e., only insider cites during selection period are used to form bins in panel A and only outsider cites in panel B. Regressions use subsample of observations from botany and ecology journals described in Table 6. Specification is otherwise identical to that in Table 4; see that table for applicable notes.

dent variable is insider cites in panel A and outsider cites in panel B. The results for insider cites exaggerate the monotonic pattern compared to any yet seen, even those interacted with PubMed access, the estimate for the 0-cite bin is the most strongly negative and for the 11+ cite bin is the most strongly positive. Outsider cites in panel B of Table 7 are negative for all except the 11+ cite bin, which is not statistically significantly different from zero at the 10% level (although the coefficient on partial open access in this bin is negative and significant).

To help visualize patterns, the last panel of Figure 5 graphs the results from Table 7 along with quadratic curves fitting the points. The curves resemble their theoretical counterparts in Figure 3. Of the two theoretical curves for insiders in Figure 3, the dashed one for ordinary insiders and the solid black one including a fraction of smart insiders, the one for ordinary insiders one appears to fit the empirical results for insiders better. The implication is that insiders may not obtain appreciably better signals of quality in advance of access than outsiders. Comparing the theoretical and empirical curves for outsiders, the estimated open-access effect for outsiders is monotonic in quality as predicted by theory but is not everywhere negative. The positive result for the highest-quality bin may be due to noise or may reflect the possibility that outsiders also face a sanction for irrelevant cites, just smaller than insiders'.

8. Conclusions

In solving some of the methodological problems associated with estimating the causal effect of open access, our previous work (McCabe and Snyder 2014) uncovered some surprising findings. While open access was found to cause a modest positive boost to cites on average, some content was found to be harmed by open access, in particular, content in the bottom half of journals in our sample of science journals. A further puzzle was that PubMed Central, a convenient repository that should facilitate easy access, was found to boost cites less than access through the narrower platform of the journal's own website.

This paper offers theoretical and empirical arguments suggesting that the previous results were not statistical flukes but predictable features of the market for academic journals. The intuition here is that open access increases the payoff from acquiring the full text of an article (at least for scholars at institutions that cannot afford the subscription fee for the closed-access journal) while leaving the payoffs of other strategies—not citing it or citing it based on superficial information—unchanged. For the articles of the lowest possible quality, citations decrease because some readers switch from superficially citing to acquiring the full text of these articles. While these readers would have cited these articles unseen, once they acquire them, they are almost certain to find that they are not worth citing. The theory thus shows that a negative open access effect for low-quality

content is not a quirk by a predictable outcome from reasonable citation behavior on the part of readers.

The theory generates a range of predictions beyond a possible negative open-access effect for some content. The predictions are readily summarized in Figure 2, showing that the open access effect is negative for the lowest-quality articles, positive for the highest-quality articles, and monotonic and concave throughout the range of article qualities. The theory predicts that improving the convenience of open access—as would be associated with placing the article on a broad platform like PubMed Central—should rotate the open-access effect so that it is even more sensitive to article quality, exaggerating the citation losses experienced by the lowest-quality content and the citation gains experienced by the highest-quality content.

The empirical results from separate regressions for bins of articles formed on the basis of cites in a pre-study period line up closely with the theoretical predictions. Articles in lower citation bins are harmed by open access (the marginal effect as low as –12.3%, significant at the 1% level); articles in higher bins benefit (the marginal effect as high as 7.7%, significant at the 1% level). With few exceptions, the open-access effect is monotonic in quality over the bins. PubMed access rotates the open-access effect, exaggerating the citation losses for low-quality content and the gains for high-quality content, just as would be predicted by a theory under the assumption that PubMed access involves lower non-pecuniary costs than access solely though the journal's own website.

We find dramatic differences between the effect of open access on cites from insiders versus outsiders to the article's field. The open-access effect is shifted downward for outsider cites compared to insider cites, with negative effects for all quality bins but the last (which is not statistically significantly different from 0 for the full-open-access variable but is statistically significantly negative for the partial-open-access variable). This comparison is consistent with outsiders engaging in more citing unseen than insiders under closed access, which theory predicts if outsiders face little sanction for irrelevant cites, amplified by low average match quality with outside content. Open access leads outsiders to move from citing unseen to full access, often determining that the low-quality content is unsuitable for citing, leading to stronger negative open-access effects for outsiders. Though the model of smart insiders, who obtain a good signal of quality even before

accessing, raised the possibility that the open-access effect can be nonmonotonic in quality for them, our empirical estimates for insider cites did not reveal such nonmonotonicities. We may not need to assume a different process generates quality signals for the two types of reader; it may be enough to assume they differ in parameters (sanction s and mean quality $\bar{q}$).

Taken together, these results suggest fairly strong "superstar" effects of open access, a new result in this literature. This substitution away from low- to high-quality articles is evidence of better matching that would appear to benefit readers as well as authors of the higher quality papers. Authors of lower quality articles, as well as lower quality journals, appear to be the net losers in the competition for reader attention. To our knowledge we are the first to suggest a mechanism through which open-access can generate winners and losers and the first to find evidence of this possibility.

Appendix A: Proofs

Before proving the propositions stated in the text, we provide two lemmas. The first lemma shows how the type spaces change with an increase in the acquisition cost. The second lemma provides expressions for the values and derivatives of the open-access effect. To streamline the proofs, we employ the shorthand notation throughout:

$$z \equiv (1+\theta)(1-\bar{q})s. \tag{A1}$$

Lemma 1. *For all $a', a'' > 0$ such that $a' < a''$,*

$$P(B_n(a')) \leq P(B_n(a'')) \tag{A2}$$
$$P(B_u(a')) \leq P(B_u(a'')) \tag{A3}$$
$$P(B_f(a')) > P(B_f(a'')). \tag{A4}$$

Proof. Assume $0 < a' < a''$. The analysis can be divided into four cases, depending on the value of z relative to a' and a''.

First suppose $z < a'$. This implies $z < a''$ since $a'' > a'$. Then (i) is the relevant case from Figure 1 whether the acquisition cost is a' or a''. We have

$$B_n(a') = B_n(a'') = \left(0, \frac{(1-\bar{q})s}{\bar{q}}\right), \tag{A5}$$

implying

$$P(B_n(a')) = P(B_n(a'')). \tag{A6}$$

Further,

$$B_f(a') \setminus B_f(a'') = \left(\frac{a'-(1-\bar{q})s}{\theta\bar{q}}, \frac{a''-(1-\bar{q})s}{\theta\bar{q}}\right). \tag{A7}$$

This is a non-degenerate subinterval of $(0,\infty)$. To see this, note the left endpoint is positive since $a' > z$, which implies $a' > (1-\bar{q})s$. The right endpoint exceeds the left since $a' < a''$. A non-degenerate subinterval of $(0,\infty)$ has positive measure since b has a continuous distribution on $(0,\infty)$. Therefore $P(B_f(a') \setminus B_f(a'')) > 0$, implying (A4). Since the sets in (3) partition the measurable type space, $P(B_n(a)) + P(B_u(a)) + P(B_f(a)) = 1$, implying

$$P(B_u(a')) = 1 - P(B_n(a')) - P(B_f(a')) < 1 - P(B_n(a'')) - P(B_f(a'')) = P(B_u(a'')), \tag{A8}$$

where the inequality follows from (A6) and (A4).

Next, suppose $z \geq a''$. This implies $z > a'$ since $a' < a''$. Then (ii) is the relevant case from Figure 1 whether the acquisition cost is a' or a''. We have

$$P(B_u(a'))) = P(B_u(a'')) = 0. \tag{A9}$$

Further,

$$B_f(a') \setminus B_f(a'') = \left(\frac{a'}{(1+\theta)\bar{q}}, \frac{a''}{(1+\theta)\bar{q}}\right), \tag{A10}$$

which is obviously a non-degenerate subinterval of $(0,\infty)$ since $0 < a' < a''$. As argued in the previous paragraph, (A4) follows. Similar to the argument behind (A8), we have

$$P(B_n(a')) = 1 - P(B_u(a')) - P(B_f(a')) < 1 - P(B_u(a'')) - P(B_f(a'')) = P(B_n(a'')). \tag{A11}$$

Next, suppose $z \in (a', a'')$. Then (ii) is the relevant case from Figure 1 when the acquisition cost is a', and (i) is the relevant case when the acquisition cost is a''. We have

$$P(B_u(a')) = 0 < P(B_u(a'')). \tag{A12}$$

Further,

$$B_n(a'') \setminus B_n(a') = \left(\frac{a'}{(1+\theta)\bar{q}}, \frac{(1-\bar{q})s}{\bar{q}} \right). \tag{A13}$$

This is a non-degenerate subinterval of $(0,\infty)$. The left endpoint is positive since $a' > 0$, and the right endpoint exceeds the left since $z > a'$. Hence

$$P(B_n(a')) < P(B_n(a'')). \tag{A14}$$

In addition,

$$B_f(a') \setminus B_f(a'') = \left(\frac{a'}{(1+\theta)\bar{q}}, \frac{a'' - (1-\bar{q})s}{\theta\bar{q}} \right). \tag{A15}$$

This is also a non-degenerate subinterval of $(0,\infty)$. The left endpoint is positive since $a' > 0$. To see that the right endpoint exceeds the left, note

$$\frac{a'}{(1+\theta)\bar{q}} \leq \frac{(1-\bar{q})s}{\bar{q}} < \frac{a'' - (1-\bar{q})s}{\theta\bar{q}}, \tag{A16}$$

where the first inequality follows from $z \geq a'$ and the second from $z < a''$. The argument above that (A4) follows from the non-degeneracy of (A7) can be used to show (A4) also follows from the non-degeneracy of (A15).

Finally, suppose $z = a'$. All the analysis from previous paragraph carries over except for the comparison of $P(B_n(a'))$ to $P(B_n(a''))$. Now,

$$B_n(a') = \left(0, \frac{a'}{(1+\theta)\bar{q}} \right) = \left(0, \frac{(1-\bar{q})s}{\bar{q}} \right) = B_n(a''), \tag{A17}$$

where the second equality follows from $z = a'$. Equation (A17) implies that (A6) holds.

We have thus verified that (A2), (A3), and (A4) hold across the cases analyzed, which were exhaustive. *Q.E.D.*

Lemma 2. *Assume* (5) *holds. Then*

$$\Delta(q, a_c, a_o) = \frac{P(B_u(a_o)) + qP(B_f(a_o)) - [P(B_u(a_c)) + qP(B_f(a_c))]}{P(B_u(a_c)) + qP(B_f(a_c))} \tag{A18}$$

$$\Delta_q(q, a_c, a_o) = \frac{P(B_u(a_c))P(B_f(a_o)) - P(B_f(a_c))P(B_u(a_o))}{[P(B_u(a_c)) + qP(B_f(a_c))]^2} \tag{A19}$$

$$\Delta_{qq}(q,a_c,a_o) = \frac{-2P(B_f(a_c))\Delta_q(q,a_c,a_o)}{P(B_u(a_c))+qP(B_f(a_c))} \tag{A20}$$

$$\Delta(0,a_c,a_o) = \frac{P(B_u(a_o))-P(B_u(a_c))}{P(B_u(a_c))} \tag{A21}$$

$$\Delta(1,a_c,a_o) = \frac{P(B_n(a_c))-P(B_n(a_o))}{1-P(B_n(a_c))}. \tag{A22}$$

Proof. Equation (A18) follows from substituting from (4) into (6), (A19) from differentiating (A18) and rearranging, and (A20) from differentiating (A19) and rearranging. Equations (A21) and (A22) follow from substituting $q = 0$ and $q = 1$, respectively, into (A18).

It remains to check that the denominators of (A18)–(A22) are non-zero, ensuring the expressions are well-defined. Given (5) holds, (i) is the relevant case from Figure 1 under closed access, implying $P(B_u(a_c)) > 0$, implying the denominators in (A18)–(A22) are positive for all $q \in [0,1]$. *Q.E.D.*

Proof of Proposition 1. Assume (5), implying $a_c > z$. Then (i) is the relevant case from Figure 1 under closed access, implying $B_u(a_c)$ and $B_f(a_c)$ are non-degenerate subintervals of $(0,\infty)$. Non-degenerate subintervals of $(0,\infty)$ have positive measure since b has a continuous distribution on $(0,\infty)$. Hence,

$$P(B_u(a_c)) > 0 \tag{A23}$$

$$P(B_f(a_c)) > 0. \tag{A24}$$

Applying Lemma 1, substituting a_o for a' and a_c for a'', yields

$$P(B_n(a_o)) \leq P(B_n(a_c)) \tag{A25}$$

$$P(B_u(a_o)) \leq P(B_u(a_c)) \tag{A26}$$

$$P(B_f(a_o)) > P(B_f(a_c)). \tag{A27}$$

Now

$$0 < P(B_u(a_c))[P(B_f(a_o))-P(B_f(a_c))] \tag{A28}$$

$$\leq P(B_u(a_c))P(B_f(a_o))-P(B_u(a_o))P(B_f(a_c)), \tag{A29}$$

where (A28) follows from (A23) and (A27) and (A29) follows from (A26). Conditions (A28)–(A29) imply that the numerator of (A19) is positive. The denominator is positive by (A23)–(A24). Hence $\Delta_q(q,a_c,a_o) > 0$. Substituting this inequality along with (A24) into (A20) yields $\Delta_{qq}(q,a_c,a_o) < 0$.

Having characterized the derivatives of $\Delta(q,a_c,a_o)$, we are left to determine its sign. The analysis is divided into three cases, depending on the value of a_o relative to z.

First, suppose $a_o > z$. The arguments in the proof of Lemma 1 leading up to (A6) can be repeated, substituting a_o for a' and a_c for a'', yielding $P(B_n(a_o)) = P(B_n(a_c))$. Substituting into (A22) yields $\Delta(1,a_c,a_o) = 0$. Now $\Delta_q(q,a_c,a_o) > 0$ and $\Delta(1,a_c,a_o) = 0$ implies $\Delta(q,a_c,a_o) < 0$ for all $q \in (0,1)$.

Next, suppose $a_o = z$. The arguments in the proof of Lemma 1 leading up to (A17) can be repeated, substituting a_o for a' and a_c for a'', yielding $P(B_n(a_o)) = P(B_n(a_c))$. Using arguments from the previous paragraph, the same result can be obtained, namely $\Delta(q, a_c, a_o) < 0$ for all $q \in (0,1)$.

Finally, suppose $a_o < z$. The arguments in the proof of Lemma 1 leading up to (A12) can be repeated, substituting a_o for a' and a_c for a'', yielding $P(B_u(a_o)) < P(B_u(a_c))$. Substituting into (A21) yields $\Delta(0, a_c, a_o) < 0$. Since $\Delta(q, a_c, a_o)$ is differentiable, it is continuous. By continuity, $\Delta(q, a_c, a_o) < 0$ for a neighborhood above $q = 0$ as well. We can also repeat the arguments in the proof of Lemma 1 leading up to (A14), again substituting a_o for a' and a_c for a'', to obtain $P(B_n(a_o)) < P(B_n(a_c))$. Substituting into (A22) yields $\Delta(1, a_c, a_o) > 0$. By continuity, $\Delta(q, a_c, a_o) > 0$ for a neighborhood below $q = 1$ as well. *Q.E.D.*

Proof of Proposition 2. Assume (5), implying $a_c > z$. The arguments in the proof of Proposition 1 showing that (A23) and (A24) hold also apply here. Applying Lemma 1, substituting a_o^B for a' and a_o^N for a'', yields

$$P(B_n(a_o^B)) \leq P(B_n(a_o^N)) \tag{A30}$$

$$P(B_u(a_o^B)) \leq P(B_u(a_o^N)) \tag{A31}$$

$$P(B_f(a_o^B)) > P(B_f(a_o^N)). \tag{A32}$$

Substituting a_o^B and a_o^N for a_o in (A19), differencing, and rearranging, yields

$$\begin{aligned} &\Delta_q(q, a_c, a_o^B) - \Delta_q(q, a_c, a_o^N) \\ &= \frac{P(B_f(a_c))\left[P(B_u(a_o^N)) - P(B_u(a_o^B))\right] + P(B_u(a_c))\left[P(B_f(a_o^B)) - P(B_f(a_o^N))\right]}{\left[P(B_u(a_c)) + qP(B_f(a_c))\right]^2}. \end{aligned} \tag{A33}$$

The first term in the numerator is non-negative by (A24) and (A31). The second term in the numerator is positive by (A23) and (A32). The denominator is positive by (A23). Thus (A33) is positive, implying $\Delta_q(q, a_c, a_o^B) > \Delta_q(q, a_c, a_o^N)$.

Substituting a_o^B and a_o^N for a_o in (A21), differencing, and rearranging, yields

$$\Delta(0, a_c, a_o^B) - \Delta(0, a_c, a_o^N) = \frac{P(B_u(a_o^B)) - P(B_u(a_o^N))}{P(B_f(a_c))}. \tag{A34}$$

The numerator is non-positive by (A31). The denominator is positive by (A24). Thus (A34) is non-positive, implying $\Delta(0, a_c, a_o^B) \leq \Delta(q, a_c, a_o^N)$.

Substituting a_o^B and a_o^N for a_o in (A22), differencing, and rearranging, yields

$$\Delta(1, a_c, a_o^B) - \Delta(1, a_c, a_o^N) = \frac{P(B_n(a_o^N)) - P(B_n(a_o^B))}{P(B_u(a_c)) + P(B_f(a_c))}. \tag{A35}$$

The numerator is non-negative by (A30). The denominator is positive by (A23) and (A24). Thus (A35) is non-negative, implying $\Delta(1, a_c, a_o^B) \geq \Delta(1, a_c, a_o^N)$. *Q.E.D.*

Proof of Proposition 4. Letting $x^S(q,a)$ and $x(q,a)$ denote the number of cites from smart and ordinary insiders, respectively, we have

$$x^I(q,a) = \sigma x^S(q,a) + (1-\sigma)x(q,a). \tag{A36}$$

Intuitively, no smart insider cites an article with $q = 0$, implying $x^S(0,a) = 0$. This intuition can be verified in Figure 1 by replacing $\bar{q}$ with q and setting $q = 0$. We then see from the figure that $B_n(a)$ occupies the whole measurable space. Substituting $x^S(0,a) = 0$ into (A36) yields $x^I(0,a) = (1-\sigma)x(0,a)$. Thus,

$$\Delta^I(0,a_c,a_o) = \frac{x^I(0,a_o) - x^I(0,a_c)}{x^I(0,a_c)} = \frac{(1-\sigma)[x(0,a_o) - x(0,a_c)]}{(1-\sigma)x(0,a_c)} = \Delta(0,a_c,a_o).$$

We next examine the other extreme of article quality, $q = 1$. Intuitively, all smart insiders cite an article with $q = 1$ since citing unseen provides a benefit with no risk of sanction, so the reader prefers this to ignoring the article. Acquiring the full text instead always generates a cite. This intuition can be verified in Figure 1 by replacing $\bar{q}$ with q and setting $q = 1$. We then see from the figure that (i) is the relevant case, and sets $B_u(a)$ and $B_f(a)$ span the whole measurable space. Hence $x^S(1,a) = P^S(B_u^S(a)) + 1 \cdot P^S(B_f^S(a)) = 1 - P^S(B_n^S(a)) = 1$, implying $x^I(1,a) = \sigma + (1-\sigma)x(1,a)$. Substituting,

$$\Delta^I(1,a_c,a_o) = \frac{\sigma + (1-\sigma)x(1,a_o) - [\sigma + (1-\sigma)x(1,a_c)}{\sigma + (1-\sigma)x(1,a_c)} = \frac{x(1,a_o) - x(1,a_c)}{\sigma/(1-\sigma) + x(1,a_c)}. \tag{A37}$$

Since $\sigma > 0$ and $\Delta(1,a_c,a_o) = [x(1,a_o) - x(1,a_c)]/x(1,a_c)$, the last claim of the proposition follows. *Q.E.D.*

Proof of Proposition 5. In the extended model in which both a and s differ across regimes, let $B_n(a_c,s_c)$ and $B_n(a_o,s_o)$ denote the set of types who do not cite under, respectively, closed and open access. Denote the other type sets B_u and B_f analogously.

Consider open-access sanctions $s_o', s_o'' > 0$ with $s_o'' > s_o'$. First, suppose (7) does not hold. Then the relevant case from Figure 1 is (ii) when the sanction cost is s_o'; it is also the relevant case when the sanction cost is the yet higher s_o''. In case (ii), the sets do not depend on s. Hence, $B_u(a_o,s_o') = B_u(a_o,s_o'')$ and $B_f(a_o,s_o') = B_f(a_o,s_o'')$, implying $x(q,a_o,s_o') = x(q,a_o,s_o'')$ by (4).

Next, suppose (7) holds. Then (ii) is the relevant case from Figure 1 when the sanction cost is s_o'. Either case (i) or (ii) may be relevant when the sanction cost is s_o''. Take the two subcases in turn. First, suppose

$$a_o > (1+\theta)(1-\bar{q})s_o'', \tag{A38}$$

implying (i) is the relevant case when the sanction cost is s_o''. We have

$$B_n(a_o,s_o') = \left(0, \frac{(1-\bar{q})s_o'}{\bar{q}}\right) \subset \left(0, \frac{(1-\bar{q})s_o''}{\bar{q}}\right) = B_n(a_o,s_o'') \tag{A39}$$

$$B_f(a_o,s_o') = \left(\frac{a_o - (1-\bar{q})s_o'}{\theta\bar{q}}, \infty\right) \subset \left(\frac{a_o - (1-\bar{q})s_o''}{\theta\bar{q}}, \infty\right) = B_f(a_o,s_o''), \tag{A40}$$

implying

$$P(B_n(a_o,s_o')) < P(B_n(a_o,s_o'')) \tag{A41}$$
$$P(B_f(a_o,s_o')) < P(B_f(a_o,s_o'')) \tag{A42}$$

since b is continuously distributed on its support $(0,\infty)$. Now,

$$x(q,a_o,s_o) = P(B_u(a_o,s_o)) + qP(B_f(a_o,s_o)) \tag{A43}$$
$$= 1 - P(B_n(a_o,s_o)) - (1-q)P(B_f(a_o,s_o)), \tag{A44}$$

where (A43) follows from (4) and (A44) from $P(B_u(a_o,s_o)) = 1 - P(B_n(a_o,s_o)) - P(B_f(a_o,s_o))$. Equations (A41)–(A44) imply $x(q,a_o,s_o') > x(q,a_o,s_o'')$.

Finally, suppose (7) holds but (A38) is violated. Then

$$B_n(a_o,s_o') = \left(0, \frac{(1-\bar{q})s_o'}{\bar{q}}\right) \subset \left(0, \frac{a_o}{(1+\theta)\bar{q}}\right) = B_n(a_o,s_o'') \tag{A45}$$

$$B_f(a_o,s_o') = \left(\frac{a_o - (1-\bar{q})s_o'}{\theta\bar{q}}, \infty\right) \subset \left(\frac{a_o}{(1+\theta)\bar{q}}, \infty\right) = B_f(a_o,s_o''). \tag{A46}$$

In both (A45) and (A46), the first equality and strict inclusion follow from (7); the last equality follows from the violation of (A38). Equations (A45)–(A46) imply (A41)–(A42), which together with (A43)–(A44) imply $x(q,a_o,s_o') > x(q,a_o,s_o'')$. *Q.E.D.*

Appendix B: Supplementary Exhibits

Table B1: Journals in Sample

Ecology		Botany		Multidisciplinary Science and Biology	
Rank	Journal	Rank	Journal	Rank	Journal
5	*Annual Rev. Ecol. Systematics*	4	*Plant Cell*	1	*Proc. National Acad. Sci.*
6	*Advances Ecol. Res.*	9	*Annual Rev. Phytopathology*	2	*Nature*
7	*Ecol. Monographs*	10	*Plant Physiology*	3	*Science*
8	*Trends Ecol. Evolution*	12	*Plant Molecular Bio.*	16	*Proc.: Bio. Sci.*
11	*Amer. Naturalist*	15	*Planta*	17	*Philosophical Trans.: Bio. Sci.*
13	*Evolution*	18	*Molecular Plant-Microbe Interact.*	49	*Amer. Scientist*
14	*Ecol.*	19	*Plant Cell Environ.*	55	*Annals New York Acad. Sci.*
20	*J. Animal Ecol.*	21	*Botanical Rev.*	56	*Naturwissenschaften*
22	*Behavioral Ecol.*	24	*Photosynthesis Res.*	58	*Comptes Rendus Acad. Sci.*
23	*J. Ecol.*	28	*Theoretical Applied Genetics*	60	*Proc. Japan Acad. Series B*
25	*Marine Ecol.*	29	*New Phytologist*	67	*Trans. Royal Soc. South Africa*
26	*Paleobiology*	32	*Plant Cell Physiology*	73	*J. Royal Soc. New Zealand*
27	*Ecol. Applications*	33	*Protoplasma*	82	*South African J. Sci.*
30	*Oecologia*	34	*J. Experimental Botany*	90	*Current Sci.*
31	*Oikos*	35	*Physiologia Plantarum*	92	*Interciencia*
37	*Microbial Ecol.*	36	*J. Phycology*	94	*Archives Sci.*
38	*J. Applied Ecol.*	39	*Amer. J. Botany*	96	*Ohio J. Sci.*
42	*J. North Amer. Benthological Soc.*	40	*Phytopathology*		
43	*Functional Ecol.*	41	*Annals Missouri Botanical Garden*		
44	*Theoretical Population Bio.*	48	*Physio. Molec. Plant Pathology*		
45	*J. Evolutionary Bio.*	51	*Systematic Botany*		
46	*J. Experimental Marine Bio. Ecol.*	62	*Int. J. Plant Sci.*		
47	*Conservation Bio.*	75	*Functional Plant Bio.*		
50	*J. Chemical Ecol.*	100	*J. Torrey Botanical Soc.*		
52	*Evolutionary Ecol.*				
53	*J. Biogeography*				
54	*Polar Bio.*				
57	*J. Wildlife Manag.*				
59	*Bio. Conservation*				
61	*Biotropica*				
63	*Sarsia*				
64	*Environ. Bio. Fishes*				
65	*New Zealand J. Ecol.*				
66	*Ecol. Modelling*				
68	*Acta Oecologica*				
69	*J. Tropical Ecol.*				
70	*Agricultural Ecosystems Environ.*				
71	*Pedobiologia*				
72	*Biochemical Systematics Ecol.*				
74	*J. Soil Water Conservation*				
76	*Amer. Midland Naturalist*				
77	*Rangeland Ecol. Manag.*				
78	*J. Arid Environ.*				
79	*J. Natural Hist.*				
80	*Wildlife Soc. Bull.*				
81	*Proc. Acad. Natural Sci. Phila.*				
83	*Population Ecol.*				
84	*J. Freshwater Ecol.*				
85	*African J. Ecol.*				
86	*Rev. Ecol.-La Terre Et La Vie*				
87	*South African J. Wildlife Res.*				
88	*Revista Chilena Hist. Natural*				
89	*Northwest Sci.*				
91	*Canadian Field-Naturalist*				
93	*Western North Amer. Naturalist*				
95	*Bull. Amer. Museum Natural Hist.*				
97	*Biocycle*				
98	*Natural Hist.*				
99	*Russian J. Ecol.*				

Notes: Classification into ecology versus botony versus general science according to ISI primary subject. Journals ranked 1-100 within our sample using ISI impact factor averaged over 1984-2004.

Table B2: Results Binning by Citation for Alternative Citation Bins

	Cites in selection period			
	0 cites	1–2 cites	3-9 cites	10+ cites
Partial open access	0.000	–0.064	–0.009	0.033***
	(0.041)	(0.046)	(0.020)	(0.012)
Full open access	–0.046	–0.089**	–0.022	0.077***
	(0.060)	(0.034)	(0.025)	(0.015)
Articles	31,008	35,212	32,986	19,589
Panel observations	162,735	177,150	160,561	95,108
Article fixed effects	Yes	Yes	Yes	Yes
Publication × citation year fixed effects	Yes	Yes	Yes	Yes
Partial online-access indicator	Yes	Yes	Yes	Yes
Full online-access indicator	Yes	Yes	Yes	Yes
Journal-specific age profile	Linear	Linear	Linear	Linear

Notes: Specification is identical to Table 4 except uses an alternative partition to for bins. Each column is a separate regression including observations for articles having the specified number of cites in selection period. Observations in selection period are omitted from the regressions, reducing the sample size relative to that reported in Table 1. Additional notes from Table 4 apply.

Table B3: Results Binning by Publication-Year Citation Percentiles

	Publication-year percentiles				
	0–50%	50–62.5%	62.5–75%	75–87.5%	87.5–100%
Partial open access	–0.025	–0.083	–0.038	–0.002	0.030**
	(0.032)	(0.052)	(0.055)	(0.020)	(0.013)
Full open access	–0.059	–0.124***	–0.062**	–0.014	0.071***
	(0.047)	(0.032)	(0.059)	(0.027)	(0.015)
Articles	41,239	18,079	19,492	19,957	20,028
Panel observations	205,940	90,840	97,978	100,239	100,557
Article fixed effects	Yes	Yes	Yes	Yes	Yes
Publication × citation year fixed effects	Yes	Yes	Yes	Yes	Yes
Partial online-access indicator	Yes	Yes	Yes	Yes	Yes
Full online-access indicator	Yes	Yes	Yes	Yes	Yes
Journal-specific age profile	Linear	Linear	Linear	Linear	Linear

Notes: Each column is a separate regression including observations for articles whose cites in the selection period (first two years after publication) fall into that percentile compared to other articles published in the same year. Observations in selection period are omitted from the regressions, reducing the sample size relative to that reported in Table 1. Additional notes from Table 4 apply.

Table B4: Insider/Outsider Results Forming Citation Bins Using All Citations

	Cites in selection period				
Variable	0 cites	1 cite	2–5 cites	6–10 cites	11+ cites
	A. Insider cites				
Partial open access	–0.151**	–0.136***	–0.007	0.025	0.105*
	(0.062)	(0.045)	(0.025)	(0.035)	(0.059)
Full open access	–0.187***	–0.150***	0.013	0.064	0.087
	(0.047)	(0.027)	(0.028)	(0.041)	(0.056)
Articles	15,715	12,695	20,027	4,522	1,756
Panel observations	83,146	64,871	98,972	21,287	8,168
	B. Outsider cites				
Partial open access	–0.021	–0.033	–0.099**	0.021	–0.122*
	(0.045)	(0.056)	(0.039)	(0.070)	(0.055)
Full open access	–0.045	–0.032	0.190***	–0.121***	–0.127**
	(0.080)	(0.054)	(0.025)	(0.041)	(0.049)
Articles	16,439	12,789	19,873	4,490	1,750
Panel observations	87,227	65,613	98,642	21,181	8,150
Article fixed effects	Yes	Yes	Yes	Yes	Yes
Publication × citation year fixed effects	Yes	Yes	Yes	Yes	Yes
Partial online-access indicator	Yes	Yes	Yes	Yes	Yes
Full online-access indicator	Yes	Yes	Yes	Yes	Yes
Journal-specific age profile	Linear	Linear	Linear	Linear	Linear

Notes: Results are comparable to Table 7 except that, instead of using just insider cites during selection period to form bins in panel A or just outsider cites in panel B, all citations during selection period are used in both panels. Specification is otherwise identical to that in Table 7; see that table for applicable notes.

References

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

Antelman, Kristin. (2004) "Do Open-Access Articles Have a Greater Research Impact?" *College & Research Libraries* 65: 372–382.

Armstrong, Mark. (2015) "Opening Access to Research," *Economic Journal* 125: F1–F30.

Atchison, Amy and Jonathan Bull. (2015) "Will Open Access Get Me Cited? An Analysis of the Efficacy of Open Access Publishing in Political Science," *Political Science & Politics* 48: 219–137.

Besancenot, Damien and Radu Vranceanu. (2017) "A Model of Scholarly Publishing with Hybrid Academic Journals," *Theory and Decision* 82: 131–150.

Bergstrom, Theodore. (2001) "Free Labor for Costly Journals?" *Journal of Economic Perspectives* 15: 183–198.

Bergstrom, Theodore and Carl T. Bergstrom. (2004) "The Costs and Benefits of Library Site Licenses to Academic Journals," *Proceedings of the National Academy of Sciences* 101: 897–902.

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

Bryan, Kevin A. and Yasin Ozcan. (2020) "The Impact of Open Access Mandates on Invention," *Review of Economics and Statistics* forthcoming.

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

Craig, Iain D., *et al.* (2007) "Do Open Access Articles Have Greater Citation Impact? A Critical Review of the Literature," *Journal of Informetrics* 1: 239–248.

Davis, Philip M., *et al.* (2008) "Open Access Publishing, Article Downloads, and Citations: Randomised Controlled Trial," *British Medical Journal* 337: 568–573.

Davis, Philip M. and Michael J. Fromerth. (2007) "Does the arXiv Lead to Higher Citations and Reduced Publisher Downloads for Mathematics Articles?" *Scientometrics* 71: 203-–215.

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.

Eger, Thomas and Marc Scheufen. (2018) *The Economics of Open Access: On the Future of Academic Publishing*. Cheltenham, U.K.: Edward Elgar Publishing.

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: 395–399.

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: 692–698.

Feess, Eberhard and Marc Scheufen. (2016) "Academic Copyright in the Publishing Game: A Contest Perspective," *European Journal of Law & Economics* 42: 263–294.

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

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

Jeon, Doh-Shin and Jean-Charles Rochet. (2010) "The Pricing of Academic Journals: A Two-Sided Market Perspective," *American Economic Journal: Microeconomics* 2: 222–255.

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

Li, Yan, Chaojiang Wu, Erjia Yan, and Kai Li. (2018) "Will Open Access Increase Journal CiteScores? An Empirical Investigation over Multiple Disciplines," *PLOS One* 13(e0201885): 1–21.

Machado, José A. F. and J. M. C. Santos Silva. (2005) "Quantiles for Counts," *Journal of the American Statistical Association* 100: 1226–1237.

McCabe, Mark J. and Christopher M. Snyder. (2005) "Open Access and Academic Journal Quality," *American Economic Review Papers & Proceedings* 95: 453–458.

McCabe, Mark J. and Christopher M. Snyder. (2007) "Academic Journal Pricing in a Digital Age: A Two-Sided-Market Model," *B.E. Journal of Economic Analysis & Policy (Contributions)* 7(issue 1, article 2): 1–37.

McCabe, Mark J. and Christopher M. Snyder. (2014) "Identifying the Effect of Open Access on Citations Using a Panel of Science Journals," *Economic Inquiry* 53: 1284–1300.

McCabe, Mark J. and Christopher M. Snyder. (2015) "Does Online Availability Increase Citations? Theory and Evidence from a Panel of Economics and Business Journals," *Review of Economics and Statistics* 97: 144–165.

McCabe, Mark J. and Christopher M. Snyder. (2018) "Open Access as a Crude Solution to a Hold-up Problem in the Two-Sided Market for Academic Journals," *Journal of Industrial Economics* 66: 301–349.

McCabe, Mark J., Christopher M. Snyder, and Anna Fagin. (2013) "Open Access versus Traditional Journal Pricing: Using a Simple 'Platform Market' Model to Understand Which Will Win (and Which Should)," *Journal of Academic Librarianship* 39: 11–19.

Mueller-Langer, Frank and Marc Scheufen. (2013) "Academic Publishing and Open Access," chapter 23 in Ruth Towse and Christian Handke, eds., *Handbook on the Digital Creative Economy*. Cheltenham, U.K.: Edward Elgar.

Mueller-Langer, Frank and Richard Watt. (2010) "Copyright and Open Access for Academic Works," *Review of Economic Research on Copyright Issues* 7: 45–65.

Mueller-Langer, Frank and Richard Watt. (2018) "How Many More Cites is a $3,000 Open Access Fee Buying You? Evidence from a Natural Experiment," *Economic Inquiry* 56: 931–954.

Piwowar, Heather, Jason Priem, Vincent Lariviére, Juan Pablo Alperin, Lisa Matthias, Bree Norlander, Ashley Farley, Jevin West, and Stefanie Haustein. (2018) "The State of OA: A Large-Scale Analysis of the Prevalence and Impact of Open Access Articles," *PeerJ* 6(e4375): 1–23.

Scheufen, Marc. (2015) "On the Access Principle in Science: A Law and Economics Analysis," chapter 4 in Marc Scheufen, ed., *Copyright Versus Open Access: On the Organization and International Political Economy of Access to Scientific Knowledge*. Cheltenham, U.K.: Springer.

Shavell, Steven. (2010) "Should Copyright of Academic Works be Abolished?" *Journal of Legal Analysis* 2: 301–358.

Staudt, Joseph. (2020) "Mandating Access: Assessing the NIH's Public Access Policy," *Economic Policy* forthcoming.

Tang, Min, James D. Bever, and Fei-Hai Yu. (2017) "Open Access Increases Citations of Papers in Ecology," *Ecosphere* 8(e01887): 1–9.

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.

Yan, Erjia and Kai Li. (2018) "Which Domains Do Open-Access Journals Do Best in? A 5-Year Longitudinal Study," *Journal of the Association for Information Science and Technology* 69: 844–856.

Young, Jeffrey R. (2004) "Google Tests Search Engine for Colleges' Scholarly Materials," *Chronicle of Higher Education*, April 23, p. A36.