Research Policy 40 (2011) 1332–1338

Contents lists available at ScienceDirect

Research Policy

journal homepage: www.elsevier.com/locate/respol

Getting cited: Does open access help?☆

Patrick Gaulé[a,*], Nicolas Maystre[b,1]

[a] Sloan School of Management, Massachusetts Institute of Technology, MA, United States
[b] United Nations Conference on Trade and Development (UNCTAD), Geneva, Switzerland

ARTICLE INFO

Article history:
Received 6 November 2010
Received in revised form 1 April 2011
Accepted 27 May 2011
Available online 2 July 2011

Keywords:
Open access
Knowledge diffusion
Scientific publishing
Citations
Self-selection

ABSTRACT

Cross-sectional studies typically find positive correlations between free availability of scientific articles ('open access') and citations. Using a number of instruments as plausible sources of exogeneous variation, we find no evidence for a causal effect of open access on citations. We provide theory and evidence suggesting that authors of higher quality papers are more likely to choose open access in hybrid journals which offer an open access option. Self-selection mechanisms may thus explain the discrepancy between the positive correlation found in Eysenbach (2006) and other cross-sectional studies and the absence of such correlation in the field experiment of Davis et al. (2008).

1. Introduction

The dominant business model in scientific publishing is 'reader pays', i.e. university libraries pay for academic journals through subscriptions. However, scientific articles are increasingly available for free ('open access') under different modalities. Three factors have facilitated the emergence of open access: sharp decreases in dissemination costs with the advent of electronic publishing, growing expectations that the results of publicly funded research should be freely available and increased strains on library budgets associated with substantial increases in journal prices (McCabe, 2002; Dewatripont et al., 2006).

The most visible form of open access has been the creation of journals that are free for readers and financed through fees levied on authors upon submission. The journals of the Public Library of Science and its flagship journals, *PLoS Biology* and *PLoS Medicine*, are perhaps the most notable examples but the directory of open access journals currently lists more than 3000 entries. Despite concerns that open access journals may be of lower quality (Jeon and Rochet., 2007; McCabe and Snyder, 2006), some have established themselves as prestigious outlets. For instance, the open access journal *PLoS Pathogens* has an impact factor above nine and is the leading journal in the field of parasitology.

Separately, publishers are increasingly offering authors the possibility to buy open access to their articles in subscription-based journals. Initially pioneered by a number of not-for-profit publishers, open access options are now offered by almost all major publishers.[2]

Free online availability can also result from authors posting versions of their papers on their websites or in institutional repositories. This type of open access has become increasingly important with the adoption of institutional mandates. For instance, the pub-

☆ We thank seminar participants at the University of Geneva, the Graduate Institute for International and Development studies, the EPFL, the CUSO doctoral workshop, the University of Munich, Bocconi University, the AEA 2010 conference and the 'Economic Perspectives on Scholarly Communication in a Digital Age' workshop at the University of Michigan for interaction and comments. We are indebted to an anonymous referee, Stephen Bazen, Annamaria Conti, Philip Davis, Mathias Dewatripont, Jaime de Melo, Chiara Franzoni, Marco Fugazza, Alfonso Gambardella, Bronwyn Hall, Jaya Krishnakumar, Rafael Lalive, Francisco Lissoni, Christopher Liu, Jacques Mairesse, Thierry Mayer, Fabio Montobbio, Mario Piacentini, Gilbert Ritschard, Jean-Charles Rochet, Dominic Rohner, Cédric Tille, Alexis Walckiers and especially Dominique Foray, Dietmar Harhoff, Bo Honoré, Marcelo Olarreaga, Frederic Robert-Nicoud and Mathias Thoenig for insightful discussions and advice.

* Corresponding author.
E-mail addresses: pgaule@mit.edu (P. Gaulé), patrickgaule@gmail.com, nicolas.maystre@unctad.org (N. Maystre).

[1] The views and opinions expressed herein are those of the author and do not necessarily reflect those of the United Nations.

[2] The Entomological Society of America and the American Society of Limnology and Oceanography were among the first to sell optional open access, beginning in 2001 (Walker, 2004). The Company of Biologists offers an open access option in its journals *Development, Journal of Cell Science, Journal of Experimental Biology* since January 2004. *Proceedings of the National Academy of Science* started to offer an open access option in May 2004. The major publishers have followed, although not for all their journals: Elsevier ('Sponsored articles'), Springer ('Open Choice'), Blackwell ('Online Open'), Taylor & Francis ('iOpen Access'), John Wiley & Sons ('Funded Access'), Oxford University Press ('Oxford open').

0048-7333/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.respol.2011.05.025

lic access policy of the National Institutes of Health (NIH) requires authors of research funded by NIH to make their papers available for free to the public on PubMed Central no later than 12 months after publication.

The emergence of open access raises an important and interesting question. Do articles that are freely available get more widely diffused as a result? Using citations as a proxy for diffusion, many cross-sectional studies have found a positive correlation between open access and citations. The seminal contribution is Lawrence (2001) who finds that computer science conference articles that were openly accessible on the Web were cited more often than those that were not (+150%). Studies that find an open access citation advantage by comparing sample means include Walker (2004); Antelman (2004); Harnad and Brody (2004), and Norris et al. (2008). A number of papers have investigated the effect on citations of depositing a (free) preprint on arXiv, an archive of working papers in mathematics and physics (Schwarz and Kennicutt, 2004; Kurtz et al., 2005; Metcalfe, 2005; Moed, 2007; Davis and Fromerth, 2007). While papers deposited on arXiv get cited more, the interpretation is disputed with the more recent studies attributing this effect to quality differentials or earlier availability. Evans and Reimer (2009) follow a different approach by comparing citations before and after scientific publications become freely available and find a small effect of open access (around 8%). This study has been criticized by McCabe and Snyder (2011) for failing to take into account time trends in citations.

Perhaps the most influential cross-sectional study is Eysenbach (2006) who compares the citation rates between open access and non-open access articles published in the second half of 2004 in *Proceedings of the National Academy of Sciences* (PNAS), a journal that offers an open access option for a fee. The strength of his approach is that he is able to control for important observables that might affect both the choice of open access and citations. Controlling for number of authors, authors' lifetime publication count and impact, submission track, country of corresponding author, funding organization, and discipline, he finds that open access articles were twice more likely to be cited in the first 4–10 months after publication. Davis (2009) reports a small but significant open access advantage for both PNAS and a number of other journals that offer an open access option.

Doubts about whether a positive correlation between open access and citations could be interpreted as a causal effect have long existed. However, such doubts acquired particular salience with recent field experimental evidence from Davis et al. (2008). This study found no differences in citations between papers randomly assigned to open access and control papers after one year (nor after three years, as shown in the follow-up study (Davis, 2011)). However, the external validity of this study is limited by the fact that the sample of journals participating in the study may not be representative of the underlying population of journals. Another issue in interpreting these field-experimental results is that potential readers did not know that they could obtain the full-text for free, unless they browsed the journal website, or explicitly searched for the article.

In this paper, we attempt to reconcile the field experimental evidence of Davis et al. (2008) with the results of cross-sectional studies, and in particular Eysenbach (2006). We first show explicitly in a simple model why comparisons of means for articles from a hybrid journal might lead to upward biased estimates of open access. A larger readership is especially valuable for the authors[3] if the paper is of high quality: for a given increase in the number of readers, a higher quality paper will receive more additional citations than a lower quality paper. Thus, open access is relatively more attractive to authors of high quality papers and thus open access papers tend to be of higher quality on average. Consequently, regressions of the number of citations on open access capture both a diffusion effect and a self-selection effect.

Empirically, we analyze a sample of 4388 biology papers published between May 2004 and March 2006 by *Proceedings of the National Academy of Sciences* (PNAS) an important, high-volume scientific journal which started to offer an open access option to authors in May 2004 for a USD 1000 fee.

In this journal, open access papers receive significantly more citations after controlling for observables, as found in Eysenbach (2006) who favored a diffusion (causal) interpretation. After replicating this cross-sectional correlation using a broader sample, we extend the analysis in multiple ways and reach opposite conclusions. We first find empirical evidence of self-selection using an original measure of article quality, i.e. the ratings from F1000 biology, a website where biology professors evaluate new papers of interest. We also implement an instrumental variable strategy where our preferred instrument is a dummy for publication of the article in the last quarter of the fiscal year. The idea here is that academic departments may have unused budgets that must be spent before the end of the fiscal year (or the funds are lost). Thus, discretionary spending on otherwise low-priority items such as paying for optional open access fees is more likely to be observed towards the end of the year, which is born out by our data. Using this instrument, we find that the coefficient of open access is insignificant and dramatically reduced compared to the coefficient of a simple ordinary least squares regression. Similar results are found with other instruments (and combinations thereof): a change of publication policy for NIH intramural researchers and a dummy for Howard Hughes Medical Institute investigators (who receive a special budget to pay open access fees).

Our results cast serious doubts on the causal interpretation of the open access advantage observed in Eysenbach (2006) and other observational studies. Instead self-selection mechanisms explain at least part of the open access citation advantage observed in such studies. Although our point estimates suggest no causal effect of open access at all, a quantitatively small causal effect cannot be statistically ruled out.

The rest of the article is organized as follows. Section 2 introduces a simple model of the open open access choice. Section 3 describes the data used in this paper. The econometric specification and results are presented in Section 4. Section 5 provides more additional evidence on self-selection versus diffusion as an explanation of the open access citation advantage. Section 6 concludes.

2. A simple model

We formalize here the idea that open access is relatively more attractive to authors of higher quality papers and its implications. This is a model of the decision to buy open access after the paper is accepted in a journal. Let q_i be the quality of the article defined as the probability of the article being cited conditional on the article being read. q_i is exogeneously given and heterogenous across articles. The number of citations N generated by an article of quality q_i is thus $N(q_i, n) = nq_i$ where n is the number of readers. Authors value citations as they help them secure peer recognition, jobs, promotions and continued research funding (Stephan, 1996). However, the present value of a citation may vary across authors for instance according to age and career stage. δ_j is the (heterogeneous and exogeneously given) present value of a citation.

[3] As researchers care about the visibility of their work, they may be willing to pay to ensure that their work receives a larger number of citations. Indeed the present value of a single additional citation for a 35-year-old physicist's work was estimated to exceed 3000 current dollars (Diamond, 1986).

Authors maximise the present value of the number of citations generated by an article minus the publication cost c:

$$U_A = \delta_j n q_i - c \quad (1)$$

Authors can choose to publish in open access (OA) or in restricted access (RA). The publication cost for the author is c_{OA} if he publishes in open access and zero otherwise. The number of readers is n_{OA} if the article is published in open access and n_{RA} otherwise, with $n_{OA} \geq n_{RA}$. Utility maximisation thus involves resolving a tradeoff between the costs of publication and a larger readership. An author will choose to publish in open access if

$$(n_{OA} - n_{RA})\delta_j q_i \geq c_{OA} \quad (2)$$

The comparative statics are straightforward: a paper is more likely to be published in open access if the quality q_i of the paper is high, if the present value of a citation δ_j is high, if the cost of publishing c_{OA} in open access is low and if the increase in readership associated with open access $(n_{OA} - n_{RA})$ is high.

The cost of purchasing open access is constant. However, the benefit to the author increases with the quality of the paper. The latter occurs because higher quality papers enjoy a larger increase in citations for a given increase in readership. Thus, in equilibrium, the average quality of open access papers is higher than that of restricted access papers.

This has important implications empirically. What we really would like to know is the percentage increase in the number of citations for an article of a given quality. However, what we observe is the percentage difference of citations between open access papers and restricted access papers. Since being in open access is not randomly assigned but is the outcome of a maximization process, the observed difference in citations is upward biased.

3. Data

3.1. The PNAS dataset

Our original dataset consists of 4388 articles published in the scientific journal *Proceedings of the National Academy of Sciences* (PNAS) between May 2004 and March 2006.[4] PNAS is an important scientific journal which is second in reputation only to *Nature*, *Science* and *Cell*. It publishes a high volume of primary research papers (weekly issues with 60 papers per issue). Restricting the analysis to a single important journal enables us to have a more homogeneous sample and to focus on within-journal variability.

Upon acceptance of their papers, PNAS authors are offered the possibility to buy open access exchange for a USD 1000 fee. If they pay the fee, the electronic version of the paper is available for free on the journal website. If they choose not to buy open access, access is restricted to subscribers for the first six months. In any case, readers based in developing countries have free and immediate access to all articles.

We focus on articles published in the area of biological sciences which represents approximatively 90% of papers published in PNAS. An important point is that contrary to economics or physics, circulation of pre-publication papers (working papers, preprints, ...) is inexistent in biology where pre-publication would significantly decrease the chances of subsequent publication in an academic journal. Self-archiving by authors is also uncommon. To verify that, we searched for full text versions of articles published in one issue of PNAS three months after its publication. Of the 43 articles published in restricted access, we were able to find only two cases where a full-text version was freely available elsewhere on the web.

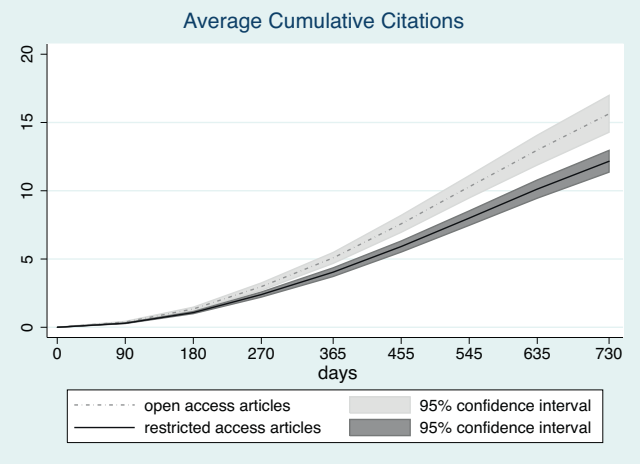

Fig. 1. Cumulative citations to PNAS articles.

For cited papers, we know from the website of the journal whether the article was published in open access or not, the names of the authors, the publication date, the subfield in which it was published, the email address of the corresponding author, the submission track[5] and whether the article was featured on the cover of the journal.

3.2. Citation data

Citation data were extracted from ISI Web of Science which is the standard bibliometric database and includes citations from over 7000 scientific journals. For citing papers, we know the time of publication and the journal where they are published. We use this information to construct the cumulative number of citations after various lengths of time.

Fig. 1 displays the mean and 95% confidence intervals of citations accumulated over time for both open access and restricted access papers. About 17% of our sample consists of open access papers. A citation advantage of open access article is apparent from the raw data.

For the rest of the paper we focus on the number of citations accumulated within two years as our dependent variable. We chose citations after two years for two reasons. First, it is a very conventional time window to observe citations, for instance the journal impact factors calculated by ISI/Thompson are based upon citations after two years. Second, this was the maximal amount of time for which we could observe citations when the data was collected. While the treatment (free availability versus restricted availability) is limited to 6 months, its effect on citations might appear later. The first reason for this is very mechanical; even though publication cycles are much shorter in biology than in economics, there is nevertheless a potentially substantial lag between the time of first submission (which we do not observe) and the time of publication. For instance, a citing article may be submitted five months after the publication of the cited article, spend 6 months in the refereeing process and wait another two months to move from accepted

[4] Our sample is larger than the original Eysenbach (2006) study and we observe citations over a longer period. However, the main difference is that in our analysis we use instruments as sources of exogeneous variation and have additional control variables.

[5] In additional to the usual submission track where authors submit manuscripts to the editorial office, this journal has two special submission tracks. Academy members can submit their own papers with two referee reports to the editorial office (track III). They can also communicate manuscripts from other authors that they find interesting (track I).

Table 1
Descriptive statistics.

	Open access ($n = 723$)		Restricted access ($n = 3665$)	
	Mean	S.D.	Mean	S.D.
Dependent variable (Y):				
Citations after two years	17.98	22.06	13.55	12.31
Control variables (X):				
Last author productivity	0.28	0.36	0.26	0.33
Number of authors	4.38	1.03	6.05	3.72
Years since 1st pub. of the last author	24.43	10.29	24.31	10.35
F1000 "grade" (*)	0.98	1.95	0.82	1.81
Broad appeal	0.03	0.18	0.02	0.12
Last author is a star	0.14	0.34	0.12	0.32
From the cover	0.09	0.29	0.07	0.26
Submission = Track II (standard submission)	0.36	0.48	0.5	0.5
Submission = Track III (academy members)	0.31	0.46	0.27	0.44
Private firms	0.06	0.24	0.02	0.15
National Institutes of Health (NIH)	0.05	0.21	0.04	0.19
Instruments (Z):				
NIH – post reform	0.04	0.19	0.02	0.13
End of fiscal year	0.21	0.41	0.16	0.37
Howard Hughes Medical Institute (HHMI)	0.13	0.34	0.06	0.24

(*) Grades on F1000biology.com are clustered around 3 ("recommended"), 6 ("must read") and 9 ("exceptional"). Papers not evaluated coded as 0.

status to actual publication. The second reason is initial citations may accelerate subsequent citations (an indirect effect).[6] Indeed, it is a common theme in the economics and sociology of science that small initial advantages can be magnified over time.

3.3. Controls

Last author characteristics: We focus on the last author because in the natural and physical sciences, a robust social norm systematically assigns last authorship to the principal investigator (Azoulay et al., 2006; Riesenberg and Lundberg, 1990). We construct two proxies to control for quality of the last author. First, we match the names of the last author with Medline data extracted using PublicationHarvester (Azoulay et al., 2006). We use these data to construct the variable 'Last author productivity' which is defined as the number of publications of the last author weighted by the impact factor of the publishing journal and divided by the number of years since (s)he started publishing.[7] Second, we construct a dummy that takes value 1 if the last author is a star, i.e. if (s)he is appears on one of ISI Web of Science lists of 250 most cited researchers in various subfields of biology. Finally, our regressions also include the number of years since the first publication as a proxy for age of the last author.

Article quality: We use a novel proxy for article quality which is the evaluation given by biology professors on the website F1000 Biology.[8] Contributors to this website post short summaries of recently published papers together with an evaluation which can be either 'recommended', 'must read' or 'exceptional'. The contributors are university professors and experts in particular subfields of biology. Around 19% of articles in our sample have received an evaluation on F1000: 12% appear as 'recommended', 6% as 'must read' and 1% as 'exceptional'. A dummy is constructed for each of these types of evaluations. While this proxy for article quality contains only four different modalities (not evaluated, recommended, must read, and exceptional), which might not capture entirely the quality of the article, it, nevertheless, yields useful insights.[9]

Since open access might be motivated by a desire to facilitate access to readers outside the scientific community, we also construct a dummy 'Broad appeal' that takes value 1 if the article was cited in *Scientific American*, *New Scientist*, *the Economist* (Table 1).

3.4. Instruments

Our empirical strategy consists of instrumenting open access to isolate the effect of diffusion from self-selection. Our preferred instrument is a dummy for publication in the last quarter of the fiscal year. We exploit here the fact that academic departments may have leftover budgets that need to be spent before the end of the fiscal year.[10] One otherwise low-priority item on which budgets can be spent is paying for open access fees for papers about to be published in PNAS. While there is evidence of fiscal year seasonality influencing economic outcomes (Oyer, 1998), to the best of our knowledge we are the first to use it as an instrument. In our data, 21% of articles published in the last quarter of the fiscal year are in open access compared to 15% for the three other quarters. At Harvard, where the fiscal year ends on the 30th of June, 42% of articles published in April, May and June are in open access compared to 15% for those published in the rest of the year.

Our second instrument is a dummy that takes value 1 if the corresponding author is an intramural researcher of the National Institutes of Health (NIH) and the article was published after April

[6] We thank an anonymous referee for this remark.

[7] One problem we encountered is that it is difficult to identify publications for authors with common last names. The procedure we used to deal with this issue was to exclude observations where the last author had a very common last name (more than 5 occurrences of different authors with the same last name in our dataset). This results in a loss of 590 observations mainly for papers with last authors with an Asian name. For moderately common names (between 2 and 5 occurrences of different authors with the same last name in our dataset), we kept them in the dataset but adjusted the total number of publications downwards by dividing the total number by the number of different occurrences in the dataset. The results of our paper are robust to alternative specifications.

[8] http://www.f1000biology.com.

[9] One might think that using our proxy as a control for quality raises two potential concerns. First, open-access may increase the likelihood that a given article receives an evaluation on F1000. We think that is highly unlikely as F1000 contributors are eminent scientists, and it is hard to imagine that they lack institutional or personal access to PNAS. Second, there could be a positive feedback loop whereby a higher number of citations increases the likelihood that a given article receives an evaluation on F1000. We think that this is highly unlikely as well, as the vast majority of evaluations are made shortly (less than 1 month) after the article is published, before citations can be observed.

[10] We coded the end of the fiscal year as follows: end of June for US-based academic institutions; end of September for US government, end of December for other countries.

Table 2
Results.

	Pooled OLS (I) Two years citations	2SLS 1st stage (II) Open access	2SLS 2nd stage (III) Two years citations	GMM 2nd stage (IV) Two years citations	LIML 2nd stage (V) Two years citations
Open access	4.123a [0.581]		0.813 [4.494]	0.363 [4.413]	0.795 [4.518]
F1000 = "recommended"	3.442a [0.628]	0.02 [0.017]	3.526a [0.657]	3.573a [0.651]	3.527a [0.657]
F1000 = "must read"	4.946a [0.813]	0.055b [0.026]	5.135a [0.865]	5.184a [0.860]	5.136a [0.866]
F1000 = "exceptional"	6.571b [2.642]	0.064 [0.084]	6.812a [2.536]	6.755a [2.534]	6.812a [2.536]
Broad appeal	2.448 [1.867]	0.118a [0.048]	2.826 [1.980]	2.900 [1.976]	2.828 [1.981]
Last author is a star	3.067a [0.781]	0.012 [0.020]	3.106a [0.791]	3.160a [0.786]	3.106a [0.791]
From the cover	6.164a [0.877]	−0.001 [0.022]	6.157a [0.879]	6.135a [0.870]	6.157a [0.879]
Last author productivity	2.709a [0.789]	0.000 [0.021]	2.746a [0.783]	2.719a [0.783]	2.746a [0.784]
Submission = Track II	−0.149 [0.405]	−0.090 a [0.014]	−0.442 [0.564]	−0.476 [0.560]	−0.443 [0.566]
Submission = Track III	−1.358 a [0.493]	−0.042 b [0.017]	−1.479 a [0.513]	−1.483 a [0.513]	−1.480 a [0.514]
Number of authors	0.616a [0.083]	−0.023 a [0.002]	0.541a [0.127]	0.531a [0.125]	0.540a [0.127]
Years since 1st pub. of the last author	−0.068 a [0.018]	−0.001 [0.001]	−0.071 a [0.018]	−0.071 a [0.018]	−0.071 a [0.018]
Private firms	0.899 [1.004]	0.219a [0.039]	1.633 [1.422]	1.735 [1.410]	1.637 [1.425]
N.I.H.	−0.672 [0.753]	−0.058 c [0.033]	−0.548 [0.750]	−0.476 [0.736]	−0.548 [0.749]
End of fiscal year		0.049a [0.016]			
NIH – post reform		0.178a [0.055]			
H.H.M.I.		0.109a [0.027]			
Year FE	yes	yes	yes	yes	yes
Subfield FE	yes	yes	yes	yes	yes
Constant (Biochemistry subfield)	7.844a [0.825]	0.269a [0.025]	8.7778a [1.477]	8.901a [1.456]	8.783a [1.483]
F test on IVs			12.13		
Hansen J stat./*P*-value			0.30/0.86		
Observations	4388	4388	4388	4388	4388
R-squared	0.126	0.09			

Notes: Robust standard errors in brackets. *c* significant at 10%; *b* at 5%; *a* at 1% Column I reports the benchmark OLS regression. The first-stage of the two-stage least squares regression with the three instruments is displayed in column II. The second stage of the two stage least square regression is displayed in column III and the results of the GMM estimation and LIM in column IV and V.

2005. The NIH issued a new policy on open access in February 2005, to be implemented in May 2005. Although this policy was primarily aimed at research funded by the NIH and conducted *extra muros*, it also had an effect on NIH intramural researchers. Before the change in policy, only 13% of articles authored by NIH intramural researchers were in open access. After the change in policy, the corresponding number was 28%. Since we control for being a NIH intramural researcher and for time trends, we expect the instrument to capture only the effect of open access. Our third instrument is a dummy that takes value 1 if one of the authors is an investigator for the Howard Hughes Medical Institute (HHMI). The HHMI provides a special budget of USD 3000 to its investigators to pay for open access fees. Since HHMI investigatorships are prestigious, it is important that we control for author quality to ensure the validity of the instrument.

4. Econometric specification and results

As a benchmark we estimate with ordinary least squares and robust standard errors:

$$Y = \delta * open\_access + X\beta + \varepsilon \quad (3)$$

where Y is the number of citations after two years and X is the complete set of control variables described in the preceding section.

We then implement the instrumental variable strategy with two-stage least squares, limited information maximum likelihood (LIML) and with GMM.[11] GMM is more efficient than two-stage least squares under conditional heteroskedasticity (Hayashi, 2000) and LIML is approximately median-unbiased for overidentified constant effects models (Angrist and Pischke, 2009). We refrain from using a nonlinear first stage such as a probit or logit, because the second stage estimates would not be consistent if the functional form of the first stage was incorrect (Angrist, 2001; Angrist and Krueger, 2001).

The results of the benchmark OLS regression are reported in the first column in Table 2. The coefficient on open access is positive and significant at the 1% confidence level. The coefficient is robust to various combinations of controls. It is also quantitatively important with 4.12 more citations (+53%) for open access articles than restricted access articles. These results are in line in terms of both significance and magnitude with those of Eysenbach (2006) and Davis (2009) with similar samples. Our next regressions investigate whether this coefficient can be interpreted as causal.

The first stage of the two-stage least squares regression with the three instruments is displayed in column II. The three instruments are significant at the 1% confidence level. The first-stage *F*-statistics is 12.13 and the Stock-Yogo (2005) test statistic rejects the null that the group of instruments is weak (the critical value of the test for three instruments at a 5% confidence level, one endogeneous regressor and a 2SLS bias of 10% is 9.08).

The first stage provides evidence of self-selection of higher quality articles into open access.[12] The coefficient on our proxies for

[11] We considered two alternative estimators, a matching approach and the Heckit procedure. However, for a matching approach to work, the selection must be made on observable characteristics. As Cameron and Trivedi (2005:871) put it "the key assumption is that unobservable variables play no role in the treatment assignment [in our case, the choice of open-access] and outcome determination [in our case, the number of citations]". We were concerned that our proxy for article quality (F1000 rankings) captures only partially the quality of the article, and that we were essentially dealing with a problem of selection on unobservables, which can be addressed by instrumental variables. The Heckman two-step estimator (or Heckit estimator) is usually used when the sample is non-randomly selected (hence when there is a selection bias). In our case, we analyze the number of citations on all articles published in PNAS between May 2004 and March 2006. Hence, we do not observe only the articles published in open access, we also include in our sample the ones that are published in "restricted access". However, since we claim that the choice of open access is not exogenous to the quality of the article (which is likely to be correlated to the number of citations), we need to adopt an IV estimator.

[12] Eysenbach (2006) reports no statistically significant differences in self-reported article quality between open access articles and other articles from an author survey. Yet, we are skeptical about the use of self-reported measures from survey data as proxies for the quality of research.

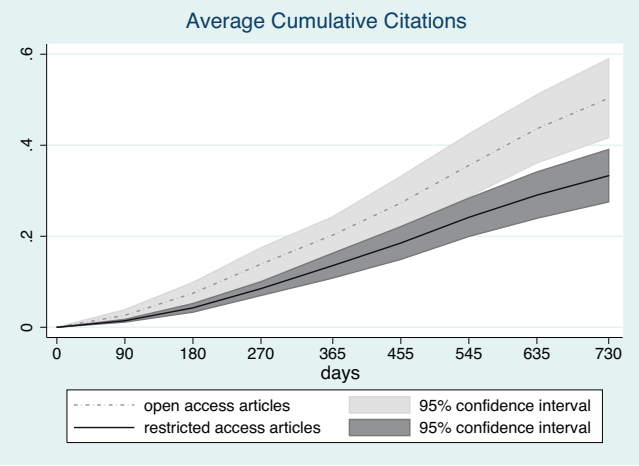

Fig. 2. Cumulative citations to PNAS articles from articles published in *Nature*, *Science* or *Cell*.

article quality (the evaluation on F1000 biology and broad appeal) are positive and significant. The dummy for 'must read' is significant at the 5% confidence level. The dummy for 'exceptional' is not significant but the number of articles in this category is very small. A joint *F*-test on the three F1000 dummies reject the null that they are not different from zero at the 5% confidence level. As robustness check, we ran a probit of open access on the same explanatory variables with the same results.

The second stage of the 2SLS is displayed in column III and the results of the GMM estimation and LIML in column IV and V respectively. When we instrument, the coefficient on open access decreases by a factor of five or more and is no longer significant. This is the case with either 2SLS, GMM or LIML.

5. More evidence on selection versus diffusion

Besides the regressions, two other arguments further suggest that the open access advantage observed in the raw data (Fig. 1) and in the non-instrumented specification (column 1) comes from self-selection rather than from a diffusion effect of open access.

First, the timing of citation accumulation over time observed in Fig. 1 seems inconsistent with a diffusion effect of open access. Given that the treatment is free versus restricted access only for the first six months (after which every article becomes freely available), one would expect the open access citation advantage to stop increasing after six-months. The opposite is observed which is not reconcile to reconcile with a diffusion or causal, effect of open access on citations. However, a causal effect cannot be ruled out from this evidence alone, because initial citations may accelerate future citations.[13]

Second, we look at citations in *Science*, *Nature* and *Cell*, the three most prestigious scientific journals. Authors publishing in these highly prestigious journals are performing cutting-edge science and can hardly be expected to lack extensive access to the scientific literature. Yet, as shown in Fig. 2, open access papers also receive significantly more citations in these three journals.

Both of these facts are at odds with a diffusion effect but can be readily explained by self-selection of higher quality articles into open access. If open access articles are on average of better quality, they should receive more attention in the top journals (hence the citation differential in *Science*, *Nature* and *Cell*) and the open access citation advantage considering all citations should continue to increase after six months.

6. Concluding remarks

The main contribution of this paper is to show that at least part of the larger number of citations received by open access papers is due to a self-selection effect rather than a diffusion (or causal) effect. We provide theory and evidence suggesting that authors of higher quality papers are more likely to choose open access in hybrid journals which offer an open access option. Self-selection mechanisms may thus explain the discrepancy between the positive correlation found in many cross-sectional studies and the absence of such correlation in the field experiment of Davis et al. (2008).

Using three instruments as plausible sources of exogeneous variation, we find no evidence for a causal effect of open access on citations. However, a quantitatively small causal effect cannot be statistically ruled out. Perhaps we should not be too surprised by the absence of a large effect. Gaule (2009) reports that biologists based in India facing important limitations in their access to the literature yet routinely obtain electronic versions of papers through requests to authors or friends who have better access.

Our results may not apply to other forms of open access beyond journals that offer an open access option. Authors increasingly self-archive either on their website or through institutional repositories. Studying the effect of that type of open access is a potentially important topic for future research.

An important limitation of studies based on citations is that they do not capture 'invisible readers', i.e. readers that do not themselves publish in scientific journals. Although the main readership of scientific papers is scientists themselves, students and practitioners occasionally read scientific articles, in particular in medicine.

The diffusion effect of open access is an interesting and important question. However, whether open access should be widely adopted ultimately depends on the sum of all its welfare effects. A full welfare analysis is beyond the scope of this paper but we note such an analysis might include the time spent by readers accessing materials and competitive effects in the scientific publishing market (Bergström, 2001; Wellcome Trust, 2003; Dewatripont et al., 2006).

[13] We thank an anonymous referee for this remark.

References

Angrist, J., 2001. Estimation of limited dependent variable models with dummy endogenous regressors: simple strategies for empirical practice. Journal of Business & Economic Statistics 19, 2–16.
Angrist, J., Pischke, J., 2009. Mostly Harmless Econometrics: An Empiricist's Companion. Princeton University Press, Princeton.
Angrist, J., Krueger, A., 2001. Instrumental variables and the search for identification: from supply and demand to natural experiments. Journal of Economic Perspectives 15, 69–85.
Antelman, K., 2004. Do open access articles have a greater research impact? College and Research Libraries 65, 372–382.
Azoulay, P., Stellman, A., Graff Zivin, J., 2006. PublicationHarvester: an open-source software tool for science policy research. Research Policy 35, 970–974.
Bergström, T., 2001. Free labour for costly journals? Journal of Economic Perspectives 15, 183–198.
Cameron, A.C., Trivedi, P.K., 2005. Microeconometrics, Methods and Applications. Cambridge University Press, New York.
Davis, P.M., Lewenstein, B., Simon, D., Booth, J., Connolly, M., 2008. Open access publishing, article downloads, and citations: randomised controlled trial. British Medical Journal 337, a568.
Davis, P.M., 2009. Author-choice open access publishing in the biological and medical literature: a citation analysis. Journal of the American Society for Information Science and Technology 60, 3–8.
Davis, P.M., 2011. Open access, readership, citations: a randomized controlled trial of scientific journal publishing. FASEB Journal 25.
Davis, P.M., Fromerth, M.J., 2007. Does the arXiv lead to higher citations and reduced publisher downloads for mathematics articles? Scientometrics 71, 203–215.
Dewatripont, M., Ginsburgh, V., Legros, P., Walckiers, A., Devroey, J.P., Dujardin, M., Vandooren, F., Dubois, P., Foncel, J., Ivaldi, M., Heusse, M.P., 2006. Study on the economic and technical evolution of the scientific publication markets in Europe. European Commission.

Diamond, A., 1986. What is a citation worth? Journal of Human Resources 21, 200–215.
Evans, J.A., Reimer, J., 2009. Open access and global participation in science. Science 323, 1025.
Eysenbach, G., 2006. Citation advantage of open access articles. PLoS Biology 4, e157.
Gaule, P., 2009. Access to the scientific literature in India. Journal of the American Society for Information Science & Technology 60, 2548–2553.
Harnad, S., Brody, T., 2004. Comparing the impact of Open Access (OA) vs. non-OA articles in the same journals. D-Lib 10 (6).
Hayashi, F., 2000. Econometrics. Princeton University Press, Princeton.
Jeon, D.S., Rochet, J.C., 2007. The Pricing of Academic Journals: A Two-Sided Market Perspective. Economics Working Paper 1025, Universitat Pompeu Fabra.
Kurtz, M.J., Eichhorn, G., Accomazzi, A., Grant, C., Demleitner, M., Henneken, E., et al., 2005. The effect of use and access on citations. In: Information Processing and Management, 41, pp. 1395–1402.
Lawrence, S., 2001. Free online availability substantially increases a paper's impact. Nature 411, 521.
McCabe, M.J., 2002. Journal pricing and mergers: a portfolio approach. American Economic Review 92, 259–269.
McCabe, M.J., Snyder, C.M., 2006. The Economics of Open Access Journals. Working paper. http://mccabe.people.si.umich.edu/EOAJ.pdf (accessed on March 1, 2010).
McCabe, M.J., Snyder, C.M., 2011. Did Online Access to Journals Change the Economics Literature? Available at SSRN: http://ssrn.com/abstract=1746243.
Metcalfe, T.S., 2005. The rise and citation impact of astro-ph in major journals. Bulletin of the American Astronomical Society 37, 555–557.
Moed, H.F., 2007. The effect of 'Open Access' upon citation impact: an analysis of ArXiv's Condensed Matter Section. Journal of the American Society for Information Science and Technology 58, 2047–2054.
Norris, M., Oppenheim, C., Rowland, F., 2008. The citation advantage of open access articles. Journal of the American Society for Information Science and Technology 59, 1963–1972.
Oyer, P., 1998. Fiscal year ends and nonlinear incentive contracts: the effect on business seasonality. The Quarterly Journal of Economics 113, 149–185.
Riesenberg, D., Lundberg, G., 1990. The order of authorship: who's on first? Journal of American Medical Association 264, 1857.
Schwarz, G.J., Kennicutt, R.C.J., 2004. Demographic and citation trends in astrophysical journal papers and preprints. Bulletin of the American Astronomical Society 36, 1654–1663.
Stephan, P., 1996. The economics of science. Journal of Economic Literature 24, 1199–1235.
Stock, J., Yogo, M., 2005. Testing for weak instruments in linear IV regression. In: Stock, J., Andrews, D. (Eds.), Identification and Inference for Economic Models: Essays in Honor of Thomas J. Rothenberg. Cambridge University Press, Cambridge.
Walker, T., 2004. Open access by the article: an idea whose time has come? Nature Web Focus, http://www.nature.com/nature/focus/accessdebate/13.html (accessed on March 1, 2010).
Wellcome Trust, 2003. Economic analysis of scientific research publishing. A report commissioned by the Wellcome Trust. http://www.wellcome.ac.uk/stellent/groups/corporatesite/policy_communications/documents/web_document/wtd003182.pdf (accessed on March 1, 2010).