GAULE, Patrick; MAYSTRE, Nicolas

Getting cited: does open access help?

November 2008

CEMI-WORKINGPAPER-2008-007

Keywords: scholarly publishing, open access, free access

JEL classification : O33, O38

Abstract

We reexamine the widely held belief that free availability of scientific articles increases the number of citations they receive. Since open access is relatively more attractive to authors of higher quality papers, regressing citations on open access and other controls yields upward-biased estimates. Using an instrumental variable approach, we find no significant effect of open access. Instead, self-selection of higher quality articles into open access explains at least part of the observed open access citation advantage.

Getting cited: does open access help?[*]

Patrick Gaulé[†][‡]& Nicolas Maystre[§]

November 12, 2008

Abstract

We reexamine the widely held belief that free availability of scientific articles increases the number of citations they receive. Since open access is relatively more attractive to authors of higher quality papers, regressing citations on open access and other controls yields upward-biased estimates. Using an instrumental variable approach, we find no significant effect of open access. Instead, self-selection of higher quality articles into open access explains at least part of the observed open access citation advantage.

[*] 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 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 Stephen Bazen, Annamaria Conti, Philip Davis, Jaime de Melo, Chiara Franzoni, Alfonso Gambardella, Bronwyn Hall, Jaya Krishnakumar, Rafael Lalive, Francisco Lissoni, Christopher Liu, Jacques Mairesse, Fabio Montobbio, Mario Piacentini, Jean-Charles Rochet, Dominic Rohner, Cédric Tille, Alexis Walckiers and especially Dominique Foray, Dietmar Harhoff, Bo Honoré, Marcelo Olarreaga and Mathias Thoenig for insightful discussions and advice.

[†] Chaire en Economie et Management de l'Innovation, Collège du Management de la Technologie, Ecole Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland, patrick.gaule@epfl.ch

[‡] Department of Economics, University of Geneva

[§] Department of Economics, University of Geneva, Uni Mail, Bd du Pont d'Arve 40, 1211 Genève 4, Switzerland nicolas.maystre@unige.ch

1 Introduction

The dominant business model in scientific publishing is ‘reader pays’, i.e., university libraries pay for academic journals through subscriptions. However, an alternative model is gaining momentum where authors pay and readers have free and immediate access (‘open access’) to the full text of scientific articles. The emergence of open access is facilitated by 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).

A large number of open access journals have been created- the directory of open access journals currently lists more than 3000 entries. 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[1].

Despite concerns that open access journals may be of lower quality (Jeon & Rochet 2007, McCabe & 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.

An important question in this context is whether (and by how much) open access increases the number of citations received by scientific papers. 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). Because the open access citation advantage underpins the willingness of authors and institutional actors to pay for open access, it is central to the dynamics of the scientific publishing market.

The mainstream opinion in the information science literature is that open access increases the number of citations received by scientific papers and that this effect is quantitatively important. 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 by Walker (2004) and Antelman (2004) also find an open access citation advantage by

[1] The Entomological Society of America and the American Society of Limnology and Oceanography were among the first to sell open access by the article, 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’).

comparing sample means of citations[2]. The most influential of these studies 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). 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.

This view has been challenged by Davis et al. (2008) who randomly allocated papers from journals of the American Physiological Society into open access and found no open access citation advantage after one year. In the appendix, we report similar results for an experiment undertaken by the *Journal of Medical Genetics*. While these results suggest that there might be no effect at all, they do not quite settle the debate. In particular, they do not provide a convincing explanation of the mechanism generating the open access advantage observed in the earlier studies. To the extent that the effect of open access is heterogeneous across journals, it could be that there is no effect in the journals where the experiments were undertaken but that there is an effect in other journals.

In this paper, we first show explicitly in a simple model why comparisons of means might lead to upward biased estimates of open access. A larger readership is especially valuable 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. We 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 reduced compared to the coefficient of a simple ordinary least squares regression. Similar results are found with other instruments (and combinations thereof): a change

[2] Walker (2004) reports that articles published by the American Society of Limnology and Oceanography with open access in 2003 were downloaded more often (+180%) than online articles accessible only to subscribers. Antelman (2004) finds that freely available articles were more frequently cited than those in restricted in electric engineering (+51%), mathematics (+91%), philosophy (+45%) and political science (+86%).

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 are consistent with the self-selection of higher quality article into open access. Although a diffusion effect of open access cannot be ruled out, our results suggest that it is smaller than previously thought, if it exists at all.

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 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. 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.

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

$$U_A = \delta_j n q_i - c \tag{1}$$

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

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

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 fact that a paper is more likely to be published in open access if the quality of the paper is high 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

We use data from a single journal which enables us to have a more homogeneous sample and to focus on within-journal variability. 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. PNAS is an important scientific journal which is second only to *Nature* and *Science* in reputation and impact factor within the multidisciplinary science journals. It publishes a high volume of primary research papers (weekly issues with 60 papers per issue).

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.

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[3] 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 includes citations from over 7000 scientific journals. For citing papers, we know the time of publication and the journal where they

[3] 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).

are published. We use this information to construct the cumulative number of citations after various lengths of time.

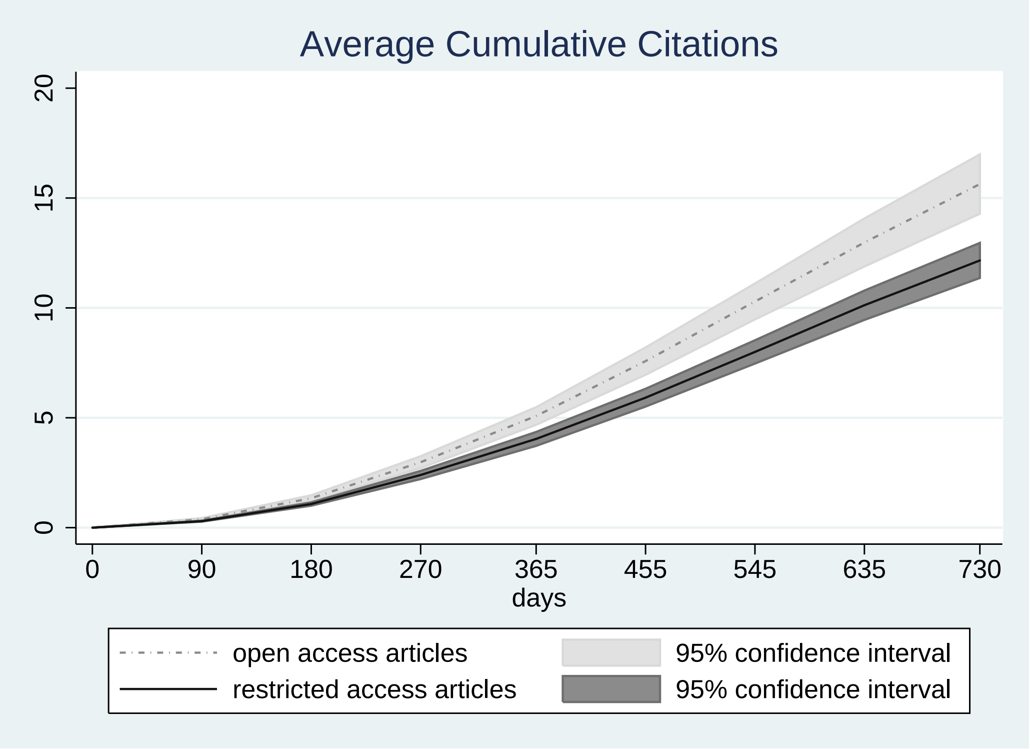

Figure 1: Mean total number of citations for open access articles (dotted line) with 95 percent confidence interval (light gray) and restricted articles access articles (solid line) accumulated after various time intervals. The sample is all research articles published in Proceedings of the National Academy of Sciences between May 2004 and March 2006 (5498 papers).

Figure 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 as this is where we have most variability.

3.3 Controls

Author quality. We construct two proxies to control for author quality. First, we match the names of the last author (who is typically the head of the lab) 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[4]. Second, we construct a dummy that takes value 1 if the last author is a superstar, 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.

Article quality. We use a novel proxy for article quality which is the evaluation given by biology professors on the website F1000 Biology[5]. 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 all 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'.

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* or the French mainstream press.

Table 1 - Descriptive Statistics

Variable:	**Obs**	**Mean**	**Std. Dev.**	**Min**	**Max**
Dependent variable (Y):					
Citations after two years	4388	13.06	11.62	0	171
Control variables (X):					
Last author productivity	4388	0.27	0.34	0	3.13
Number of authors	4388	5.77	3.48	1	60
Years since 1^{st} publication	4388	24.31	10.36	0	36
Open access	4388	0.17	-	0	1
F1000 = "recommended"	4388	0.12	-	0	1
F1000 = "must read"	4388	0.06	-	0	1
F1000 = "exceptional"	4388	0.01	-	0	1
Broad appeal	4388	0.02	-	0	1
Superstar	4388	0.12	-	0	1
From the cover	4388	0.08	-	0	1
Submission = Track II (standard submission)	4388	0.48	-	0	1
Submission = Track III (academy members)	4388	0.28	-	0	1
Private firms	4388	0.03	-	0	1
National Institutes of Health (NIH)	4388	0.04	-	0	1
Instruments (Z):					
NIH - post reform	4388	0.02	-	0	1
End of fiscal year	4388	0.17	-	0	1
Howard Hughes Medical Institute (HHMI)	4388	0.06	-	0	1

[4] 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.

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

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[6]. 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. The interest of this instrument is not so much its strength but the high likelihood that it is exogeneous since we expect that the timing of publication within the year to have no relationship with article quality.

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 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 an 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 the 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:

$$Y = \delta * open\_access + X\beta + \varepsilon \tag{3}$$

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

[6] 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.

We then implement the instrumental variable strategy with two-stage least squares (2SLS) and with GMM which is more efficient 2SLS under conditional heteroskedasticity (Hayashi 2000). 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 & Krueger 2001).

The results of the benchmark OLS regression are reported in the first column in table 1. 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 4.091 more citations for open access articles than restricted access articles (95% confidence interval: [2.95; 5.22]). These results are in line with those of Eysenbach (2006) who concluded that open access facilitates the dissemination of research findings.

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 provides evidence of self-selection of higher quality articles into open access. The coefficient on our proxies for 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 1% 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 two-stage least squares is displayed in column III and the results of the GMM estimation in column VI. When we instrument, the coefficient on open access decreases considerably and is and no longer significant. As robustness check we run two stage least squares with different combinations of the instrument and find coefficients for open access taking values between -2.978 and 2.309 (cf. appendix).

A number of other arguments further suggest that the open access advantage observed in the raw data (figure 1) and in the non-instrumented specification (column 1) does not come from a diffusion effect of open access. In particular, PNAS papers are freely accessible to developing countries from the date of publication and PNAS is one of the least expensive scientific journals in terms of both price per article and price per citation. Open access articles enjoy an even larger citation advantage considering citations in *Science*, *Nature* and *Cell*, although authors publishing in these highly prestigious journals can hardly be expected to lack extensive access to the scientific literature. We report the details of these results elsewhere (cf. Gaule & Maystre 2008).

Table 2 - Results

	Pooled	2SLS	2SLS	GMM
	OLS	1^{st}stage	2^{nd}stage	
	(I)	(II)	(III)	(IV)
	Two years	Open	Two years	Two years
	citations	access	citations	citations
Open access	4.091a		0.857	0.423
	[0.579]		[4.482]	[4.401]
F1000 = "recommended"	3.445a	0.02	3.528a	3.573a
	[0.628]	[0.017]	[0.657]	[0.651]
F1000 = "must read"	4.933a	0.054b	5.114a	5.161a
	[0.814]	[0.026]	[0.657]	[0.651]
F1000 = "exceptional"	6.578b	0.064	6.812a	6.753a
	[2.642]	[0.084]	[2.536]	[2.564]
Broad appeal	2.847c	0.147a	3.309c	3.396c
	[1.591]	[0.044]	[1.763]	[1.755]
Superstar	3.041a	0.011	3.075a	3.126a
	[0.782]	[0.020]	[0.791]	[0.786]
From the cover	6.150a	-0.002	6.141a	6.118a
	[0.877]	[0.022]	[0.879]	[0.870]
Last author productivity	2.730a	0.001	2.769a	2.744a
	[0.789]	[0.021]	[0.785]	[0.783]
Submission = Track II	-0.137	-0.089a	-0.421	-0.453
	[0.405]	[0.014]	[0.561]	[0.557]
Submission = Track III	-1.342a	-0.042b	-1.458a	-1.461a
	[0.493]	[0.017]	[0.512]	[0.512]
Number of authors	0.615a	-0.023a	0.541a	0.532a
	[0.083]	[0.002]	[0.127]	[0.126]
Years since 1^{st} publication	-0.068a	-0.001	-0.071a	-0.071a
	[0.018]	[0.001]	[0.018]	[0.018]
Private firms	0.924	0.220a	1.643	1.742
	[1.003]	[0.040]	[1.423]	[1.410]
N.I.H.	-0.676	-0.059c	-0.556	-0.487
	[0.753]	[0.033]	[0.750]	[0.737]
End of fiscal year		0.050a		
		[0.016]		
NIH - post reform		0.180a		
		[0.055]		
H.H.M.I.		0.109a		
		[0.027]		
Year FE	yes	yes	yes	yes
Subfield FE	yes	yes	yes	yes
Constant	7.825a	0.267a	8.733a	8.850a
(Biochemistry subfield)	[0.825]	[0.025]	[1.469]	[1.451]
F test on IVs			12.13	
Hansen J stat. / P-value			0.28 / 0.87	
Observations	4388	4388	4388	4388
R-squared	0.13	0.09		

Notes: Robust standard errors in brackets. *c* significant at 10%; *b* at 5%; *a* at 1%

5 Concluding remarks

The specific contribution of this paper is to show (1) 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 effect and (2) that the diffusion effect of open access is smaller than previously thought. While the evidence presented in this paper is not favorable to a large effect of open access on citations, a small effect cannot be ruled out.

One puzzle is that in the absence of a citation advantage, we should not observe open access in equilibrium in our model. This inconsistency is more apparent than real however. If authors (or even a fraction of authors) *believe* that open access increases the number of citations, some of them will choose open access and self-selection dynamics will occur. While this belief is not rational in the simple setup of the model, it is consistent with the information set that authors had when they made the decision to buy open access- that open access increases the number of citations was a commonly held view.

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.

It is also important to keep in mind that even if open access does not have any diffusion effect, it might be welfare-enhancing for other reasons. For instance, it might save readers time by making access to the full text more convenient and avoiding the need to navigate a complicated web of digital rights management restrictions. Open access may facilitate indexing and referencing by robots (such as Google indexing) which makes scientific information easier to find. Moreover, open access may have a procompetitive effect in the scientific publishing market where anticompetitive concerns have repeatedly been voiced (Bergstrom 2001, Wellcome Trust 2003, Dewatripont et al. 2006). The successful launch of the Public Library of Science and its open access journals *PLOS Medicine* and *PLOS Biology* in 2006 was followed by profitability warnings regarding Elsevier and other important publishers.

A final thought is that it is perhaps not surprising that open access does not appear to have an important effect on citations. In a world of open science full disclosure of results is a central norm (David 2003). One manifestation of the norm is that authors are almost always willing to send electronic copies of their papers to anyone who requests them. Our advice to authors is to abide by that norm and to use the 1000-3000 dollars open access publication fee for something else, unless they have time-limited budgets to use!

6 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(1):2-16

Angrist J & Krueger A (2001) ”Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments” *Journal of Economic Perspectives* 15(4):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 ”PublicationHarvester: An Open-Source Software Tool for Science Policy Research”. *Research Policy*, 35(7), pp. 970-974, 2006

Bergström T (2001) ”Free Labour for Costly Journals?” *Journal of Economic Perspectives* 15(4):183-198

David P (2003) ”The Economic Logic of “Open Science” and the Balance between Private Property Rights and the Public Domain in Scientific Data and Information: A Primer” SIEPR Discussion Paper No. 02-30

Davis P, Lewenstein B, Simon D, Booth J & Connolly M (2008) ”Open access publishing, article downloads, and citations: randomised controlled trial” *British Medical Journal* 337:a568

Dewatripont M, Ginsburgh V, Legros P, Walckiers A, Devroey JP, Dujardin M, Vandooren F, Dubois P, Foncel J, Ivaldi M & Heusse MP (2006) ”Study on the economic and technical evolution of the scientific publication markets in Europe” Final report commissioned by DG-Research, European commission

Diamond A (1986) ”What is a Citation Worth?” *Journal of Human Resources*, 21 (2): 200-15.

Eysenbach G (2006) “Citation Advantage of Open Access Articles” *PLoS Biology* 4(5):e157

Gaule P & Maystre N (2008) ”Revisiting the open access citation advantage”

Hayashi F (2000) *Econometrics.* Princeton: Princeton University Press

Jeon DS & Rochet JC (2007) “The Pricing of Academic Journals: A Two-Sided Market Perspective” Economics Working Paper 1025, Universitat Pompeu Fabra

Lawrence S (2001) ”Free online availability substantially increases a paper’s impact”. *Nature* 411:521

Maher ER (2005) The Journal of medical genetics and open access publishing: to choose or not to choose? *Journal of Medical Genetics* 42:97

McCabe M (2002) “Journal Pricing and Mergers: A Portfolio Approach” *American Economic Review* 92(1):259-269

McCabe M & Snyder C (2005) “Open Access and Academic Journal Quality” *American Economic Review Papers and Proceedings* 95(2)

McCabe M & Snyder C (2006) “The Economics of Open Access Journals” Working paper

Oyer P (1998) ”Fiscal Year Ends And Nonlinear Incentive Contracts: The Effect On Business Seasonality,” *The Quarterly Journal of Economics* 113(1):149-185

Stephan P (1996) ”The Economics of Science” *Journal of Economic Literature* 24:1199-1235.

Walker T (2004) Open access by the article: an idea whose time has come? *Nature* web focus

Wellcome Trust (2003) “Economic analysis of scientific research publishing” A report commissioned by the Wellcome Trust

A Appendix: the Journal of Medical Genetics experiment

In 2005 and 2006, the *Journal of Medical Genetics* did an experiment whereby papers from UK authors were randomized into open access and closed access (Maher 2005). Specifically, ”a randomisation table was constructed by an independent statistician and is administered by the journal editorial assistant...” (Folkes personal communication 2008). We identified the research papers published in open access under this scheme (19 papers) and papers which would have been eligible but were published in restricted access (104 papers) and compared their citation rates and extracted all citations to papers from both groups. It turns out that the mean number of citations for open access papers is no larger than for restricted access papers (cf. figure 2).

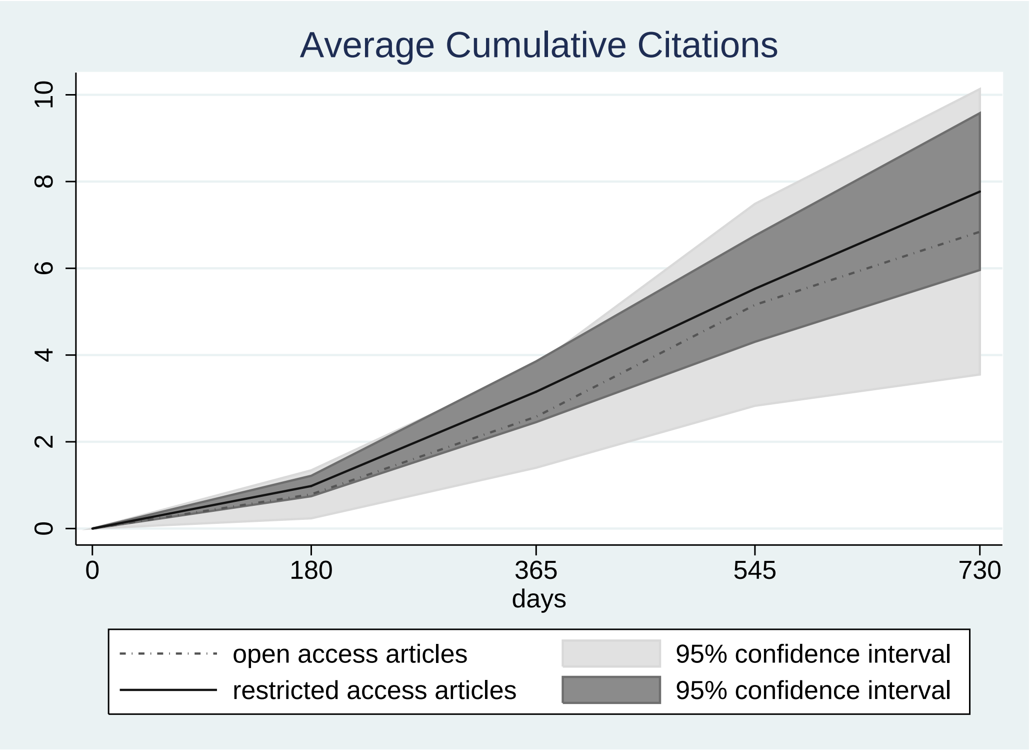

Figure 2: Mean number of citations for open access articles (dotted line with 95 percent confidence interval in light grey) and restricted access articles (solid line with 95 percent confidence in dark grey) are represented after various length of time. The confidence interval is obtained by regressing the cumulative number of citations on open access using robust standard errors. The sample is all papers published in Journal of Medical Genetics (JMG) by UK-based authors between January 2005 and December 2006.

B Robustness checks

Table 3 - Robustness check: one instrument only

	2SLS	2SLS	2SLS	2SLS	2SLS	2SLS
	1^{st}stage	2^{nd}stage	1^{st}stage	2^{nd}stage	1^{st}stage	2^{nd}stage
	(I)	(II)	(III)	(IV)	(V)	(VI)
	Open access	Two years citations	Open access	Two years citations	Open access	Two years citations
Open access		2.309		-2.978		1.912
		[9.894]		[7.913]		[6.471]
F1000 = ”recommended”	0.026	3.490a	0.025	3.626a		3.501a
	[0.017]	[0.702]	[0.017]	[0.672]	[0.017]	[0.669]
F1000 = ”must read”	0.056b	5.033a	0.057b	5.329a	0.053b	5.055a
	[0.026]	[1.004]	[0.026]	[0.939]	[0.026]	[0.903]
F1000 = ”exceptional”	0.071	6.707b	0.072	7.088a	0.066	6.736a
	[0.087]	[2.627]	[0.086]	[2.564]	[0.083]	[2.570]
Broad appeal	0.144a	3.101	0.143a	3.857c	0.145a	3.158c
	[0.044]	[2.132]	[0.044]	[2.009]	[0.044]	[1.904]
Superstar	0.011	3.060a	0.009	3.116a	0.012	3.064a
	[0.021]	[0.796]	[0.020]	[0.800]	[0.020]	[0.793]
From the cover	-0.003	6.145a	-0.002	6.129a	-0.003	6.144a
	[0.022]	[0.880]	[0.022]	[0.892]	[0.022]	[0.876]
Last author productivity	0.013	2.751a	0.013	2.815a	-0.001	2.756a
	[0.013]	[2.751]	[0.013]	[2.815a]	[-0.001]	[2.756a]
Submission = Track II	-0.089a	-0.293	-0.088a	-0.758	-0.088a	-0.328
	[0.014]	[0.960]	[0.014]	[0.815]	[0.014]	[0.687]
Submission = Track III	-0.034b	-1.406b	-0.036b	-1.596a	-0.043b	-1.420a
	[0.017]	[0.588]	[0.017]	[0.585]	[0.017]	[0.537]
Number of authors	-0.023a	0.574b	-0.023a	0.453b	-0.023a	0.565a
	[0.002]	[0.239]	[0.002]	[0.196]	[0.002]	[0.166]
Years since 1^{st} publication	-0.001	-0.069	-0.001	-0.074a	-0.001	-0.070a
	[0.001]	[0.020]	[0.001]	[0.020]	[0.001]	[0.019]
Private firms	0.218a	1.32	0.222a	2.496	0.225a	1.408
	[0.040]	[2.429]	[0.040]	[2.046[]	[0.040]	[1.753]
N.I.H.	0.037	-0.61	-0.065b	-0.414	0.043	-0.595
	[0.030]	[0.845]	[0.033]	[0.761]	[0.030]	[0.772]
End of fiscal year	0.048a					
	[0.016]					
NIH - post reform			0.181a			
			[0.055]			
H.H.M.I.					0.106a	
					[0.028]	
Year FE	yes	yes	yes	yes	yes	yes
Subfield FE	yes	yes	yes	yes	yes	yes
Constant	0.273a	8.326a	0.285a	9.810a	0.271a	8.437
(Biochemistry subfield)	[0.025]	[2.874]	[0.025]	[2.355]	[0.025]	[1.967]
F test on IVs	9.377		10.746		14.82	
Observations	4388	4388	4388	4388	4388	4388
R-squared	0.09		0.09		0.09	

Notes: Robust standard errors in brackets. *c* significant at 10%; *b* significant at 5%; *a* significant at 1%