GAULE, Patrick

Access to the scientific literature in India

February 2009

CEMI-WORKINGPAPER-2009-004

Keywords : open access, scientific publishing, developing countries, access to knowledge

JEL classification : O38

Abstract

This paper uses an evidence-based approach to assess the difficulties faced by developing country scientists in accessing the scientific literature. I compare backward citations patterns of Swiss and Indian scientists in a database of 43'150 scientific papers published by scientists from either country in 2007. Controlling for fields and quality with citing journal fixed effects, I find that Indian scientists (1) have shorter references lists (2) are more likely to cite articles from open access journals and (3) are less likely to cite articles from expensive journals. The magnitude of the effects is small which can be explained by informal file sharing practices among scientists.

Access to the scientific literature in India*

Patrick Gaulé†‡

February 23, 2009

Abstract

This paper uses an evidence-based approach to assess the difficulties faced by developing country scientists in accessing the scientific literature. I compare backward citations patterns of Swiss and Indian scientists in a database of 43'150 scientific papers published by scientists from either country in 2007. Controlling for fields and quality with citing journal fixed effects, I find that Indian scientists (1) have shorter references lists (2) are more likely to cite articles from open access journals and (3) are less likely to cite articles from expensive journals. The magnitude of the effects is small which can be explained by informal file sharing practices among scientists.

*I thank seminar participants at the CUSO doctoral workshop and the CEMI research day for interaction and comments. I am indebted to Robin Cowan, Bronwyn Hall, Nicolas Maystre, Mario Piacentini, Marcelo Olarreaga, Ed Steinmueller, Mathias Thoenig and especially Philip Davis and Dominique Foray for insightful discussions and advice. I am grateful to survey participants for their time and encouragements.

†Chaire en Economie et Management de l'Innovation, Ecole Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland, patrick.gaule@epfl.ch

‡Department of Economics, University of Geneva

1 Introduction

The scientific publishing sector has been characterized by dramatic price increases in the last three decades (Dewatripont et al. 2006; McCabe, Nevo & Rubinfeld 2006). Journals subscriptions are a significant burden on institutional libraries even in rich countries and present a major obstacle to the timely and comprehensive sharing and use of scientific information (The Wellcome Trust, 2003).

In this paper, I am looking at the access to the scientific literature in a resource-constrained environment, India, by comparison with Switzerland which has some of the best funded universities in the world.

The core of the paper is based upon citation data. I compare the backward citations patterns of Indian and Swiss scientists using a database of 43'150 scientific papers published by scientists from either country in 2007. Controlling for fields and quality with citing journal fixed effects, I find that Indian scientists (1) have shorter references lists (2) are more likely to cite articles from open access journals and (3) are less likely to cite articles from expensive journals. Moreover, the difference in the length of the reference list is more pronounced in biology and medicine where circulation of (free) pre-prints and conference proceedings is inexistent.

I complement this evidence with two other types of data. The first is library subscriptions data for the Indian Institute of Science, a leading university in India. Even in such a prestigious institution, researchers lack institutional access to one third of the 100 most important biology journals. The second is data from an online survey of Indian biologists. The survey reveals considerable heterogeneity within Indian scientists in terms of access to the literature. It also points at the importance of informal file sharing practices among scientists. In the three months prior to the administration of the survey, 84 % of respondents had either contacted an author or a friend with better access to request a copy.

There is a large literature on the effect of open access on citations and some have results pertaining to developing countries. Nicholas et al. (2007) find that the passage of the journal *Nucleic Acids Research* to an open access model in January 2005 led to an increase in usage from users located in Eastern European countries of about 20%. Evans & Reimer (2009) analyze the effect of online availability and free online availability on citations. For each paper they compute the per capita GNI for the poorest country hosting an author. They find that open access lowers the per capita GNI of the poorest country hosting citing papers. However, the effect they observe could simply be due to an increase over time of the relative share of papers authored by developing country scientists.

Closest to this paper is Frandsen's (2009) comparison of citation patterns in biology between authors from developing countries and developed countries. Frandsen concludes that authors from developing countries do not cite open access more than than authors from developed countries. However, the confidence interval for the parameter of interest is large and a positive effect cannot be statistically ruled out.

The rest of the paper is organized as follows. Section 2 presents evidence from library subscriptions data. Section 3 describes the survey of Indian biologists and its results. Section 4 details the analysis of citation data and section 5 concludes.

2 Evidence from library subscriptions data

The list of journals subscriptions is publicly available for certain Indian universities and in particular the Indian Institute of Science (IISc). Thus, I am able use this information to assess directly the access to scientific journals that researchers from the IISc have through their institution digital resources. I defined a list of the 100 most important journals in biology by taking into account both the impact factors of the journals and the number of scientific articles published yearly. The IISc library has subscriptions to 59 of the journals in this list. 36 of these subscriptions are direct subscriptions. The other 23 are through INDEST, a library consortium which has made package deals with the publishers Elsevier and Springer. Furthermore, five of the 100 most important journals in biology are in open access[1]. Thus, researchers from the IISc do not have access through their library to the remaining 36 of the 100 most important biology journals. These 36 journals include several prestigious journals from the Nature Publishing Group (e.g. *Nature Medicine*, *Nature Genetics*, *Nature Reviews Drug Discovery*) as well as *Current Biology* and *Molecular Cell*.

The Indian Institute of Science is a leading university in India. It is the institution with the most 2007 scientific publications listed in ISI Web of Science and is one of only two Indian universities appearing in the Shanghai ranking of the 500 best world universities. The library website reports spending on periodicals of 110 million rupees (USD 2.79 million). Thus, the access enjoyed by IISc researchers should be seen as an upper bound rather than representative of the situation in India.

While this approach of looking directly at subscriptions is intuitively appealing, it suffers from three important limitations. First, it is difficult to obtain the relevant information for all universities. Second, I do not know which journals researchers actually need. It could be that what is relevant to researchers is the top 10 journals in their field of specialization (say bioinformatics) rather than

[1] *PLOS Biology*, *Nucleic Acids Research*, *Journal of Experimental Botany*, *BMC Genomics*, *BMC Bioinformatics*

generalist journals in biology. Third, the absence of institutional subscriptions does not imply lack of access as access to articles can be gained by other means. For instance, scientists can buy personal subscriptions, buy articles individually or contact a librarian to initiate an inter-library loan. Alternatively, they can attempt to find a free version of the article on the web in institutional repositories or on author websites. Yet another option is to contact the author of the paper or a friend with better access to the literature.

3 Evidence from survey data

In order to gain a broader understanding of the reality in the field, I contacted the India-based corresponding authors of biology papers published in 2007 and listed in ISI Web of Science and asked them to fill a short online survey. I received 348 answers out of the target population of 2212 authors or a response rate of 15.68%[2]. In addition, more than 100 respondents left comments. Respondent bias is a concern because authors who had worse access may have been more likely to answer. Consequently, the results should be interpreted with caution but they are nonetheless interesting.

Respondents were asked to self-describe[3] the percentage of journals they had access to through their library (by choosing one in five categories).

There appears to be substantial heterogeneity in journal access. 52 respondents (15%) report having access to 90% or more of the journals they need. Conversely, 142 respondents (41%) report having access to less than 30% of the journals they need. This heterogeneity can also be seen in the comments left by respondents. Many respondents wrote that they had no problem in accessing the scientific literature at all. However, many others complained stating that they had substantial problems in accessing the literature.

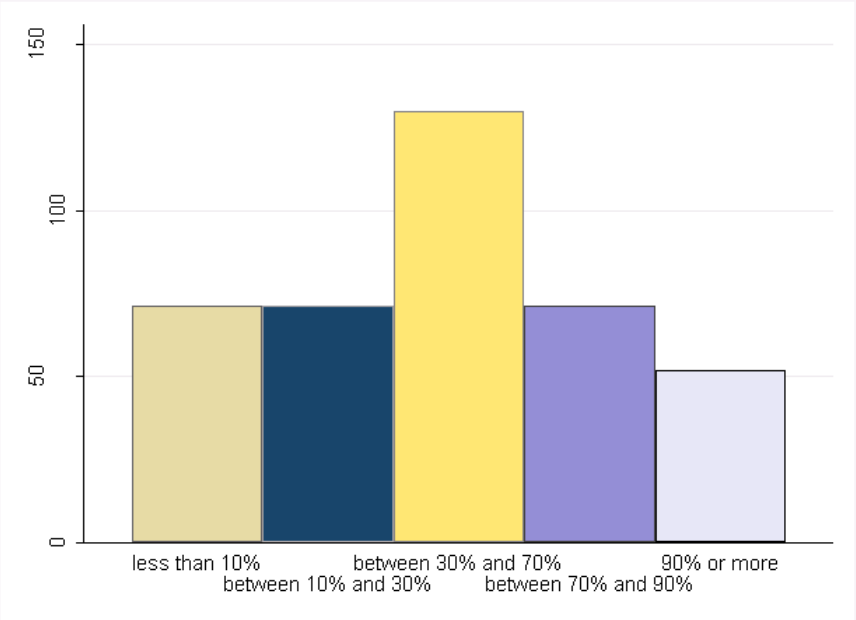
Because respondents were able to publish in journals listed in ISI Web of Science (otherwise they would not have been in the target population), I expected that most had a fairly good coverage. Even with the caveat of the possible respondents bias, it is surprising that a substantial portion of respondents report having such limited institutional access to the literature.

The survey of biologists also included a number of questions on informal files exchanges practices such as contacting the author of the paper or asking a friend with better access to the literature to

[2]For web only surveys, response rates above 10 % are considered to be good. I took no response for a no and I did not send out reminders.

[3]In the survey I have nine respondents from the IISc. Four report having access to between 30% and 70% of the journals they need, four answered between 70% and 90% and one answered more then 90%. Averaging across respondents using the middle of the intervals yields a self-reported level of 61.5% which is not far off from 64% of the library subscription evidence

Figure 1: **Journal access as self-reported by biologists in India**



In response to the question: "Regarding the access to scientific journals provided by your institution, which statement describes your situation best?", respondents could choose one of five answers: "I have access to 90% or more of the journals I need", "I have access to between 70% and 90% of the journals I need", etc.

send them a copy. Out of 332 respondents, 58.7% had sent a reprint request to an author in the last three months. Request to friends were even more frequent: 70.5% of respondents had asked a friend to send them a copy of an article in the last 3 months. Only 16% had not requested an article to either authors or friends in the last 3 months. Several respondents mentioned that they obtained articles through former students now doing research in the United States or Europe.

"In the last three months, have you...		requested a copy of an article from an author?"		
		No	Yes	
asked a friend to send you a copy of a paper that you could not access yourself?"	No	16%	25%	41%
	Yes	13.5%	45.5%	59%
		29.5%	70.5 %	100 %

Sample: 377 biologists based in India

I also asked whether the last request to an author was successful. Out of 196 respondents who had requested a copy to an author, 163 (82.7 %) had received the paper as a result of their reprint

request. Respondents had strong expectations that a corresponding author should send a copy of his paper when requested. For instance, one respondent commented that delays in responding to reprint requests were 'against the ethics of science'. Several respondents complained that these requests were sometimes (or often) ignored. One respondent noted that requests for research articles were more likely to receive an answer than requests for review articles. Several respondents pointed out that authors based in developing countries were more likely to answer positively. I also interviewed biology professors at the Universities of Geneva and Lausanne on the subject of personal reprint requests. Although reprint requests were sometimes perceived as a minor annoyance, they were nevertheless honored. Overall, the evidence suggests that reprint requests to authors are successful in most cases.

4 Evidence from citation data

4.1 Hypotheses

The idea of using citation data is that differences in access to the literature should translate into differences in citation patterns. In a rich country such as Switzerland with extensive institution-wide subscriptions most scientific articles are just 'one click away'. However in a poorer country such as India, obtaining the full text of scientific articles may involve substantial search costs (physicals visits to the library, requests to authors, librarians and friends etc.). Certain articles which would have been cited in the rich country may go unnoticed and/or unread in the poor country because the effort of finding them or getting them is higher than the expected benefit. Thus, I expect that in India the number of references cited will be lower than in Switzerland. Moreover, I expect that articles published in expensive journals will be under-represented in the citations of Indian authors relative to Swiss authors. Conversely, articles published in open access journals would be over-represented in the composition of backward citations. I further expect these effects to be moderated for Indian authors who are collaborating with authors based in OECD countries or who are based in the Indian Institutes of Technology or the Indian Institute of Science (elite institutions which receive larger budgets). Table 3.1 summarizes these hypotheses.

Table 3.1 - Expected signs

	Number references	Share expensive	Share OA
India	-	-	+
IIT	+	+	-
Collaboration_OECD	+	+	-

4.2 Data

Publication and citation data. I extracted from ISI Web of Science all 2007 publications by authors based in Switzerland and India in Science and Engineering for a total of 43150 scientific articles. I only have articles and no reviews or letters in the sample. From ISI Web of Science, I know the names of the authors, their affiliation and the journal where they published. Moreover, I know every backward citation that has been made in each of the citing articles. I have a total of 1.27 million citations. For about 91.6% of citations I know the journal in which the cited paper was published. The remaining 8.4% are citations to books, conference proceedings, scientific journals not listed in ISI Web of Science, PhD theses, etc.

Journal data. My database of information on journals comes from three different databases:

- ISI Journal Citation Report has information on the volume of articles published and the number of citations received as well as the impact factor of journals.
- Ulrich's Periodical Directory has information on journal prices and a list of open access journals. Manual checks with other sources such as the Directory of Open Access Journals (http://www.doaj.org/) and journal websites led to the conclusion that Ulrich's data is reliable and more accurate than the Directory of Open Access Journals.
- Data from www.journalprices.com by McAfee & Bergstrom. These two economists have compiled the information on journal prices from Ulrich in a useable form and made the results freely available as an information source for librarians.

Descriptive statistics. Table 3.1 displays descriptive statistics on the dependent and independent variables. 'Share expensive' is the number of references to papers published in expensive journals divided by the total number of references. An expensive journal was defined as a journal whose price per article exceeds the mean price per article in the journal database (i.e. 14.52 USD per article).[4] 'Share OA' is the number of references to papers published in open access journals divided by the total number of references. 'Collaboration_OECD' is a dummy that takes the value 1 if one of the authors is based in India and another is based in a country member of the OECD. 'IIT' is a dummy that take the value 1 is one the authors is based at the Indian Institute of Sciences or one of the Indian Institutes of Technology.

[4]The definition of an expensive journal is problematic because libraries typically buy bundles of journals at a discount over the prices listed for individual journals. However, I only observe the listed prices. Thus, I am implicitly assuming that journals with high listed prices remain more expensive that journals with low listed prices once group purchasing discounts are taken into account.

Table 3.2 - Descriptive Statistics

	India (n=27431)				Switzerland (n=15590)			
Variable:	$\boldsymbol{\mu}$	$\boldsymbol{\sigma}$	**Min**	**Max**	$\boldsymbol{\mu}$	$\boldsymbol{\sigma}$	**Min**	**Max**
Dependent variables (Y):								
Number of references	26.65	18.868	1	270	34.615	20.194	1	289
Share expensive	0.104	0.118	0	1	0.109	0.11	0	1
Share open access	0.0163	0.042	0	1	0.0136	0.0347	0	0.68
Independent variables (X):								
India	1	-	-	-	0	-	-	-
IIT	0.118	-	0	1	0	-	-	-
collaboration_OECD	0.0619	-	0	1	0	-	-	-
Number of authors	3.805	2.651	1	21	5.889	4.166	1	21

Considering these raw numbers, the reference lists of Indian authors are 22% shorter. Reference lists from Swiss authors include slightly more references to expensive journals (10.9% compared to 10.4%) and slightly less references to open access journals (1.36% compared to 1.63%).

4.3 Econometric methodology

I estimate the following regressions separately using ordinary least squares (OLS), citing journal fixed effects and robust standard errors.

$$
\begin{aligned}
number\_references_i &= \alpha_i + \beta_{1i} * India + \beta_{2i} * IIT + +\beta_{3i} * coll\_OECD + \beta_{4i} * \#authors + \epsilon_i \\
share\_open\_access_i &= \alpha_i + \beta_{1i} * India + \beta_{2i} * IIT + +\beta_{3i} * coll\_OECD + \beta_{4i} * \#authors + \epsilon_i \\
share\_expensive_i &= \alpha_i + \beta_{1i} * India + \beta_{2i} * IIT + +\beta_{3i} * coll\_OECD + \beta_{4i} * \#authors + \epsilon_i
\end{aligned}
$$

The citing journal fixed effects play an important role in the identification. By including them in the regression, I am restricting the analysis to the effect of the variables of interest on the variability of dependent variables within journals. Any differences between journals will be captured in the fixed effects. Thus, I am indirectly controlling for the quality of the journal, the field in which the journal is specialized and any other journal specific factors.

Besides the usual desirable asymptotic properties of OLS, the choice of OLS is also driven driven by computational reasons as there are more than 4000 fixed effects[5].

[5]The estimation is run in Stata using the procedure xtreg which does not actually compute the fixed effects themselves.

4.4 Results

The results of the estimations are presented in table 3.3. All variables of interest have the expected signs. The dummy on collaboration with OECD is not significant. However the other variables of interest, India and IIT are significant at the 1% confidence level in all three regressions. The reference list of an author based in India has on average 1.86 less items than the reference list of a Swiss author who publishes in the same journals. The effect is not particularly large as it corresponds to a reference list approximatively 6% shorter. This is less than in the raw data where Indian authors had on average reference lists 22% shorter. Thus, differences between journals account for most (about 3/4th) of the observed difference in the number of references. Nevertheless, there is a significant effect within journals. Papers from authors based at one of the Indian Institutes of Technology have longer reference lists than article from other India-based authors.

Table 3.3 - Results

	OLS (I)	OLS (II)	OLS (III)
	Number references	Share expensive	Share OA
India	-1.862a	-0.0091a	0.0056a
	[0.217]	[0.00145]	[0.0005]
IIT	0.817a	0.0096a	-0.0028a
	[0.298]	[0.002]	[0.0076]
Collaboration_OECD	0.553	0.0042c	-0.0011
	[0.385]	[0.00152]	[0.0058]
# authors	0.101a	-0.00063a	0.0003a
	[0.026]	[0.00018]	[0.00006]
constant	30.094a	0.114a	0.0107a
	[0.203]	[0.00136]	[0.0005]
Citing journal fixed effects	yes	yes	yes
Observations	43150	43021	43201
R-squared	0.05	0.0081	0.0049

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

Articles from open access journals represent a larger share of citations for Indian authors rather than for Swiss authors, although in either case citations to open access journals represent less than 2% of the total number of backward citations. Multiplying the coefficients from column I and III, the average article from a Swiss author has 0.32 reference to articles published in open access journals while the average article from an Indian author has 0.48 such references. Conversely articles from expensive journals represent a smaller share of citations. The effects of collaboration with OECD-based authors is weaker than expected. All these results are robust to excluding observations with

a very large number of references and to taking the log of references instead of the number of references as dependent variable.

There are important differences between fields. In the appendix I report the results of regressions estimated separately by field for the specification with the number of references. The difference between Swiss authors and Indian authors in terms of the length of the reference list is much larger in biology and medicine than in physics, engineering and chemistry. This is hardly surprising as the latter fields are characterized by a greater importance of (free) preprints and conference proceedings as means of scholarly communication[6].

Overall, the results show systematic differences between Indian and Swiss authors in their citing behavior. The most plausible explanation for these differences is Indian scientists having worse access to the literature.

5 Concluding thoughts

The different types of evidence used in this paper strongly suggest that there is a problem of access to the scientific literature in India. Many researchers self-report having limited access to the literature. Objective data from library subscriptions shows that even in an elite institution such as the Indian Institute of Science, researchers lack institutional access to one third of the top 100 biology journals.

However, the most convincing evidence comes from citation data, which is both objective and exhaustive. In biology and medicine where scholarly communication happens mainly through journal articles rather than preprints and conference proceedings, the reference lists of Indian scientists are 8.9% (biology) and 10.8% (medicine) shorter than the reference lists of Swiss scientists when they publish in the same journal. The reference lists of Swiss scientists have a higher proportion of citations from expensive journals and a lower proportion of citations from open access journals than the reference lists of Indian scientists.

Assessing whether differences in citing behavior reflects a severe problem is difficult. The meaning of a citation is disputed as citations can serve multiple functions (Davis 2009). Moravcsik and Murugesan (1975) suggested that up to 40% of references in a sample of 30 papers from the *Physical Review* were redundant in the sense that a reference is made to several papers, each of which makes the same point. It could be that the 'missing references' in the lists of Indian scientists correspond to redundant references in the lists of Swiss scientists.

[6]The preprint repository ArXiv had more than 500'000 free preprints as of October 2008, mainly in the field of physics.

In any case, an important factor limiting the severity of the access problem is the prevalence of informal file sharing practices among scientists. 84 % of Indian biologists who answered my survey had either contacted an author or a friend with better access to request a copy in the last three months. The majority of requests to authors are successful. Thus, in practice, the importance of openness as a norm of science lessens the effect of restrictions imposed by publishers on access to the literature.

An important point is that access to the literature is better viewed as a continuum rather than a yes/no variable. As Davis et al. (2008) put it, "the rhetorical dichotomy of 'open' access compared with 'closed' access does not recognise the degree of sharing that takes place among an informal network of authors, libraries, and readers". I would contend that if enough effort and time is spent in physical visits to libraries or contacting authors and friends, the vast majority of the literature is ultimately accessible to Indian scientists. However, that is not quite the same as having most articles 'one click away'.

Finally, the extent to which the results of this study can be generalized to other developing countries is unclear. However, the methodology used in this paper could easily be used as a blueprint for studies with other developing countries.

References

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

Davis P (2009) "Reward or persuasion? The battle to define the meaning of a citation" *Learned Publishing* 21:5-11

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

Evans J & Reimer J (2009) "Open access and Global Participation in Science" *Science* 323:1025

Frandsen T (2009) "Attracted to open access journals: a bibliometric author analysis in the field of biology" *Journal of Documentation* 65(1):58-82

McCabe M, Nevo A & Rubinfeld D (2006) "The Pricing of Academic Journals" Berkeley program in law and economics, working paper series, paper 199

Moravcsik M & Murugesan P (1975) "Some results on the function and quality of citations" *Social Studies of Science* 5(1):86-92

Nicholas D, Huntington P & Jamali H (2007). "The impact of open access publishing (and other access initiatives) on use and users of digital scholarly journals" *Learned Publishing* 20:11-15

The Wellcome Trust (2003) "Economic analysis of scientific research publishing." A report commissioned by the Wellcome Trust.

A Questionnaire

1. Regarding the access to scientific journals provided by your institution, which statement describes your situation best?
   - I have access to 90% or more of the journals I need
   - I have access to between 70% and 90% of the journals I need
   - I have access to between 30% and 70% of the journals I need
   - I have access to between 10% and 30% of the journals I need
   - I have access to less than 10% of the journals I need
2. In the last three months, have you requested a copy of an article from an author?
   - Yes
   - No
3. Concerning your last request, did you receive a copy of the paper?[7]
   - Yes
   - No
4. In the last three months, have you RECEIVED a request for one of your papers?
   - Yes
   - No
5. Regarding the last request that you received, did you send a copy of your paper as a result?[8]
   - Yes
   - No
6. In the last three months, have you ASKED a friend to send you a copy of a paper that you could not access yourself?
   - Yes
   - No
7. In the last three months, have you BEEN ASKED by a friend to send a copy of a paper s/he could not access himself/herself?
   - Yes
   - No
8. If you have any question or comments, please enter them here.

[7] Respondents did not receive this question if they had answered no to question 2.
[8] Respondents did not receive this question if they had answered no to question 4.

B Regressions by field

Table 3.4 - Results by field

	All fields (I)	Biology (II)	Medicine (III)	Physics (IV)	Engineering (V)	Chemistry (VI)
	Number references	Number references	Number references	Number references	Number references	Number references
India	-1.862a	-3.27a	-3.17a	-0.673	-0.705	0.028
	[0.217]	[0.603]	[0.473]	[0.463]	[0.49]	[0.552]
IIT	0.817a	0.708	5.465b	1.125	0.483	0.459
	[0.298]	[1.107]	[2.622]	[0.697]	[0.453]	[0.618]
Collaboration_OECD	0.553	0.114	0.461	1.57	1.391c	-0.26
	[0.385]	[0.62]	[0.485]	[0.518]	[0.54]	[0.622]
# authors	0.101a	0.087	0.2a	-0.025	0.163	0.0053
	[0.026]	[0.082]	[0.057]	[0.043]	[0.069]	[0.093]
constant	30.094a	39.072a	29.44a	28.62a	23.4a	30.46a
	[0.203]	[0.586]	[0.435]	[0.433]	[0.487]	[0.628]
Citing journal fixed effects	yes	yes	yes	yes	yes	yes
Observations	43150	5898	8941	7536	6835	7518
R-squared	0.05	0.067	0.079	0.028	0.0179	0.0007

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