Language and Scientific Publication Statistics: a Note

(Forthcoming in *Language Problems & Language Planning*)

Bo Sandelin (Bo.Sandelin@economics.gu.se) and *Nikias Sarafoglou* (nikias.sarafoglou@mh.se)

Abstract: The number of contributions to scientific journals by authors from various OECD countries in 1998-2000 according to the commonly used ISI databases SSCI, A&HCI, and SCI Expanded is in focus. The number of contributions per million inhabitants is related to the main language of the country, the gross domestic product per capita and whether a country is a former socialist country or not. The social sciences, the arts and humanities, and the natural sciences are studied separately. It turns out that there is a tendency for a higher publication rate for English-language countries, slightly lower for countries with small languages, and even smaller for countries with large non-English languages. This is consistent with the hypotheses that there is a bias in the data bases from the ISI such that English-language journals tend to be overrepresented, that scholars from English-language countries write almost exclusively in English, and that scholars from other countries tend to publish less in English and more in their domestic language the larger is their domestic language. This calls for caution in using these databases for international comparisons of research activity.

Keywords: language; scientific publication; ISI

JEL-codes: Z00

9 pages, October 1, 2003

Language and Scientific Publication Statistics: a Note

(Forthcoming in *Language Problems & Language Planning*)

by Bo Sandelin, Department of Economics, University of Göteborg
and
Nikias Sarafoglou, Mid-Sweden University, Sundsvall

The number of publications or of citations in scholarly journals are often used as an indication of the rank in a learned community of an individual scholar, a scientific institute, or a country. Since the Institute for Scientific Information (ISI) was founded in the USA in 1961, and especially since computers made it easier to store and analyse data, this way of measuring scientific rank has exploded. Price (1986, p. 138) observed that "most people concerned with the measurement of scientific outputs have been reduced to using counts of papers, the men who write them, and the discoveries contained in them". Such counting lies partly behind statements like: "Today, academic economics is strongly dominated by North American scholars which is reflected by publications, citations and also by Nobel Prizes" (Frey and Frey, 1995). Another example is found in an evaluation of Swedish economic research, where the authors found that "if we relate the number of articles to the population of each country, Sweden is far more productive than France or West Germany. During the period 1973-89 Sweden produced about 150 articles per million inhabitants between the ages of 25 and 65, while the corresponding figure for West Germany was about 40 and for France about 30" (Persson, Stern, and Gunnarsson, 1992).

The most commonly used databases for such studies of publications and citations are the Social Sciences Citation Index (SSCI), the Arts and Humanities Citation Index (A&HCI), and the Science Citation Index Expanded (SCI Expanded), all of which are made by the ISI. In those indices, articles proper and other material in about 8600 journals are registered, as well as references in those articles.

In this note, using the above-mentioned sources, we will introduce the national language as one of the explanatory variables when analysing publications by authors from various OECD countries. We will find that English-language countries tend to have many publications per million inhabitants, but otherwise there is a tendency that the larger is the dominant language of a country, the fewer will be the number of publications per million inhabitants. This tendency remains when we control for GDP per capita and for the fact whether a country is a former socialist country or not. The result is consistent with the combined hypotheses that English-language journals are overrepresented in the databases; that scholars in English-language countries write almost only in English; and that scholars in countries with other large languages write a lot in their own language, while scholars from countries with small languages tend to write a large proportion of their scientific articles in English. These results imply that the international comparisons of the level of scientific production in various countries based on ISI indices may be misleading.

Scientific articles by author's country

Not all scientific journals are covered in the ISI database. The SSCI includes the articles in about 1700 social science journals plus selected articles from about 3300 science and technology journals; the A&HCI includes the articles in about 1130 arts and humanities journals plus selected articles from about 7000 science and social science journals; and the SCI Expanded includes material in about 5900 science and tehnical journals.[1]

Journals are added to and deleted from the database as often as every two weeks. "ISI's editorial staff reviews nearly 2000 new journal titles annually, but only10-12% of the journals evaluated is selected."

What are the criteria for including a journal? "The journal's basic publishing standards, its editorial content, the international diversity of its authorship, and the citation data associated with it are all considered." The following is emphasized in bold letters: "**English language article titles, abstracts, and keywords are essential."** Not only that, but "English language cited references are also recommended" for those journals that wish to be included.

It is explicitely stated that "ISI seeks to cover the best regional journals as well". However, even in those cases "English language bibliographic elements remain essential". Journals in other languages that do not provide information in English are excluded.

Those circumstances lead inevitably to the hypothesis that English-language journals are overrepresented in the ISI indices.

[1] Information about included journals and the selection process is given at http://www.isinet.com. The quotations in the next paragraphs are from this source.

Table 1. Publication in scientific journals during1998-2000 by author's country of residence, articles per million inhabitants (rank in parenthesis), and millions of people in the world having the country's largest language as native language.

	Social sciences		Arts and Humanities		Natural sciences		Language size
UK	924	(1)	458	(1)	3347	(12)	400
Australia	720	(2)	192	(6)	3213	(13)	400
Canada	674	(3)	328	(3)	3759	(8)	400
New Zealand	666	(4)	226	(5)	3592	(9)	400
Netherlands	533	(5)	108	(7)	4105	(4)	21
Sweden	528	(6)	62	(17)	5546	(2)	9
Iceland	504	(7)	89	(10)	3859	(6)	0.3
USA	495	(8)	308	(4)	1129	(23)	400
Finland	488	(9)	62	(16)	5332	(3)	5
Norway	480	(10)	76	(15)	3529	(10)	4.4
Denmark	463	(11)	86	(11)	3788	(7)	5.3
Switzerland	365	(12)	101	(8)	6640	(1)	91
Ireland	331	(13)	339	(2)	3864	(5)	400
Belgium	230	(14)	83	(13)	3372	(11)	21
Germany	182	(15)	86	(12)	1843	(17)	91
Austria	179	(16)	81	(14)	2989	(14)	91
France	129	(17)	90	(9)	1805	(18)	67
Luxembourg	98	(18)	16	(24)	767	(28)	91
Spain	89	(19)	43	(19)	1867	(16)	400
Czech Rep.	84	(20)	30	(20)	1287	(21)	10
Italy	77	(21)	27	(21)	1901	(15)	55
Slovak Rep.	75	(22)	46	(18)	1105	(24)	4.5
Greece	70	(23)	22	(22)	1507	(19)	12.5
Hungary	46	(24)	18	(23)	1354	(20)	15
Portugal	37	(25)	8	(26)	972	(25)	200
Japan	37	(26)	5	(28)	1254	(22)	126
Mexico	25	(27)	9	(25)	206	(30)	400
S. Korea	20	(28)	1	(30)	862	(26)	70
Poland	13	(29)	6	(27)	780	(27)	50
Turkey	10	(30)	1	(29)	267	(29)	60

The country of the author's university or institute is indicated in the database, and we have utilized this information in table 1, which, i.a., shows the number of articles per million inhabitants during 1998-2000 in scientific journals included in the SSCI, A&HCI and SCI Expanded.[2] If there are two authors from the same country, the publication is counted as one. If an article has two authors from different countries, it is counted as one article for each country.

The 30 countries in table 1 are those belonging to the OECD. They are ordered by rank of publications in social science journals. We see that there is a clear pattern for the social sciences. There are 6 mainly English-language countries in the table (United Kingdom, Australia, Canada, New Zealand, USA, and Ireland), and they occupy the top 4 places, while all of them are found on the upper half of the table.

Thus, the English-language countries are strongest, but what comes next? Not German- or French-language countries with a glorious history of learning, but countries with smaller languages like the Netherlands and the Scandinavian countries.

Non-English-language countries with larger langauges, like Germany, France, and Spain, are found further down the list.

The dominance of the English-language countries is even more manifest in the arts and humanities. Here the 6 English-language countries take the first 6 places, but the tendency for small-language countries to dominate non-English large-language countries is weaker.

The tendencies mentioned above are weakest for the natural sciences but, as we will see later, the estimated coefficients in a regression analysis have the same signs for all three areas. One reason why the relationships are not so clear for the natural sciences may be that they have since long been more universal and less nationally oriented (cf. Persson, Stern and Gunnarson, 1992).

The numerical relationships

How does the reasoning above appear in a simple statistical analysis? In this section we have estimated two regression equations for each area of research. Each country was treated as one observation, which means that we have 30 observations. The dependent variable is the one presented in table 1, i.e., the number of scientific articles per million inhabitants. (We will, however, use logarithmic values.)

We use the following independent or explanatory variables:

[2] Articles proper dominate, but the the numbers in table 1 include also other contributions such as book reviews, corrections, and contributions to duscussions. When we talk about "articles", we use the word in this broader sense. "Natural sciences" in table 1 refers to articles in journals included in the SCI Expanded.

ENGLISH is a binary variable whose value is 1 for those countries where English is the largest domestic language (Australia, Canada, Ireland, New Zealand, UK, USA) and 0 for other countries.

LANGUAGE-SIZE is the number (millions) of people in the world that have the largest domestic language of the country as their native language according to the encyclopedia *Nationalencyklopedin.*

GDP is an index for the volume of the Gross Domestic Product per capita 1998 where the value for the whole OECD-area is 100. (Source: OECD, *Main Economic Indicators*, October 2002.) Our hypothesis is that the higher the GDP, *ceteris paribus,* the more articles are published.

SOCIALIST is in principle a binary variable whose value is 1 for former socialist countries (Czech Republic, Slovak Republic, Hungary, and Poland; Germany has the value 0.2), and 0 for the other countries. This variable is included in one version of the equations since, especially in the social sciences, research followed another tradition in the socialist countries than in the rest of the OECD-countries, and some effect of this may remain.

We will take the natural logarithms of the continuous variables. This renders a neat meaning to the regression coefficients whose estimated values are found in table 2. The estimated coefficients for equation 1 thus refer to the following equation:

ln(Articles in social sciences per million inhabitants) = Constant + *a*ENGLISH +*b*lnLANGUAGE-SIZE + e

The estimates of Constant, *a*, and *b* are, as indicated above, found in table 2 together with the other estimates; e is an error term.

Table 2. Estimated regression coefficients (OLS). (t-ratios in parenthesis.) Dependent variable: (natural logarithm of) number of articles per million inhabitants.

	Social sciences		Arts and Humanities		Natural sciences	
	Eq. 1	Eq. 2	Eq. 3	Eq. 4	Eq. 5	Eq. 6
Constant	6.0 (13.2)	-1.2 (0.8)	4.4 (8.4)	-4.6 (2.0)	8.3 (23.7)	3.1 (2.1)
ENGLISH	2.7 (5.0)	2.1 (5.7)	3.2 (5.1)	2.5 (4.7)	1.2 (2.9)	0.75 (2.2)
lnLANGUAGE-SIZE	-0.39 (3.2)	-0.33 (4.0)	-0.32 (2.3)	-0.19 (1.6)	-0.26 (2.8)	-0.20 (2.6)
lnGDP		1.6 (4.9)		1.9 (4.1)		1.1 (3.7)
SOCIALIST		-0.36 (0.9)		-0.66 (1.1)		-0.01 (0.02)
Adjusted R-square	0.45	0.78	0.46	0.66	0.27	0.52
Number of observations	30	30	30	30	30	30

All the estimated coefficients in table 2 have the expected sign, although not all of them are significantly different from zero if the 30 countries are considered as a sample of an infinitely large population. Let us first look att the social sciences.

Equation 1, with only two explanatory variables, explains 45 per cent of the variance of the dependent variable. The predicted number of articles per million inhabitants increases by 2.7 times the original amount, i.e., is 270 per cent more, if a country is ENGLISH-language than if it is not, *ceteris paribus.*

Having controlled for the influence of the ENGLISH-language binary variable, the predicted number of articles per million inhabitants decreases by 0.39 per cent for each per cent increase in the number of people in the world who have the largest LANGUAGE of the country as their native language.

Equation 2 differs from equation 1 in that we have also introduced the variables GDP and SOCIALIST. We see that the estimated magnitude and significance level of the coefficients of ENGLISH and LANGUAGE-SIZE do not change dramatically as a consequence of this extension, but they change somewhat which indicates that the explanatory variables are not quite uncorrelated with each other. The estimated effect of one per cent larger GDP per capita is 1.6 per cent more articles per million inhabitants.

The t-ratio for SOCIALIST is very low, indicating that a significant isolated effect from this variable cannot be confirmed.

For the arts and humanities, the estimated coefficients of ENGLISH is 3.2 in equation 3, which means that an English-language country is predicted to have 320 per cent more articles per million inhabitants than a non-English-language country when we controll for LANGUAGE-SIZE. This prediction changes somewhat when we also controll for GDP and SOCIALIST in equation 4. The estimated influence of the magnitude of the LANGUAGE is negative in equation 3. It remains so, but to a weaker degree, when we add GDP and SOCIALIST to the equation.

The signs of the estimated coefficients for the natural sciences (equations 5 and 6) are the same as for the social sciences and the arts and humanities, although the t-ratios differ.

The results for the three areas covered in this study do not contradict the results in an earlier study of articles in economics (Sandelin and Sarfoglou, 1997, Sandelin, Sarafoglou and Veiderpass, 2000).

Conclusions

As we indicated in the introduction, the results are consistent with the following set of combined hypotheses: 1) There is a bias in the ISI databases SSCI, A&HI, and SCI Expanded such that English-language journals of a certain quality tend to be included easier than journals in other languages of the same quality. 2) Scholars in English-language countries write virtually only in English and therefore get a large share of their publications registered in the indices, which conduces to a large number of articles per million inhabitants. 3) In order to get their research known outside a very small domestic group, scholars in countries with small domestic languages also write a lot in English, but not as much as those who have English as their native language. However, 4) the larger the non-English domestic language is, the larger is the propensity to publish in that language, and the less is the propensity to publish in English; thus, the larger the non-English domestic language, the less the number of articles per million inhabitants recorded in the ISI databases tends to be. According to our estimates, the latter decreases by 0.19-0.39 per cent for each per cent of increase in LANGUAGE-SIZE.

Therefore one should be careful with rankings of countries with respect to scholarly production based on publication frequency according to the ISI datasets. Such a ranking may wholly or largely simply reflect a language-bias, even if other hypotheses might also be consistent with the findings.[3]

REFERENCES

Frey, René L. and Bruno S. Frey. 1995. Is There a European Economics? *Kyklos* 48/2: 185-86.

Persson, Olle, Peter Stern and Elving Gunnarsson. 1992. Swedish Economics on the International Scene. In Lars Engwall (ed.), *Economics in Sweden: An Evaluation of Swedish Research in Economics*. London: Routledge.

Price, Derek J. de Solla. 1986 [1963]. *Little Science, Big Science and Beyond.* New York: Columbia University Press.

Sandelin, Bo and Nikias Sarafoglou. 1997. Artikelpublicering - vad säger amerikanska databaser? *Ekonomisk Debatt* 25/3: 155-160.

Sandelin, Bo, Nikias Sarfoglou and Ann Veiderpass. 2000. The Post-1945 development of Economics and Economists in Sweden. In A.W. Coats (ed.), *The Development of Economics in Western Europe since 1945*. London: Routledge, 42-66.

[3] There are, of course, other hypotheses that would also be consistent with part of the statistical findings. If there is a tendency among scholars in English-language countries to publish a larger proportion in journals and a smaller proportion in books than is the case in other countries, then the sign of the coefficient of ENGLISH could be positive even without a language bias in the ISI material. However, the negative sign of the coefficient of LANGUAGE-SIZE would remain to be explained.