Jointly published by Kluwer Academic Publishers, Dordrecht and Akadémiai Kiadó, Budapest

Scientometrics, Vol. 51, No. 1 (2001) 335–346

Language biases in the coverage of the *Science Citation Index* and its consequences for international comparisons of national research performance

THED N. VAN LEEUWEN, HENK F. MOED, ROBERT J. W. TIJSSEN, MARTIJN S. VISSER, ANTHONY F. J. VAN RAAN

Centre for Science & Technology Studies (CWTS), Leiden University, Leiden (The Netherlands)

Empirical evidence presented in this paper shows that the utmost care must be taken in interpreting bibliometric data in a comparative evaluation of national research systems. From the results of recent studies, the authors conclude that the value of impact indicators of research activities at the level of an institution or a country strongly depend upon whether one includes or excludes research publications in SCI covered journals written in other languages than in English. Additional material was gathered to show the distribution of SCI papers among publication languages. Finally, the authors make suggestions for further research on how to deal with this type of problems in future national research performance studies.

Introduction

Due to decreasing budgets, public accountability and the drive for efficiency in the research system, the interest in bibliometric evaluation of scientific research has strongly increased in recent years. Evaluation focusing on the comparison of national R&D systems has led to influential publications in journals like *Nature* and *Science*. In his paper of February 1997 in *Science*, Sir Robert *May* compares those national systems, thereby using various types of quantitative data.[1] Scientific publication output and impact per capita for a large number of countries were presented. Important policy relevant conclusions were drawn from his tables, which caused commotion, in circles of science policy makers as well as amongst the scientific community. The cause for concern was particularly related to pitfalls/limitations in the underlying bibliometric data. Especially for Germany and also France, the findings and conclusions of the paper were rather dramatic, both countries occupying only modest positions in the rankings.

0138–9130/2001/US $ 15.00

One remedy for this apparent unbalance involves taking into account the problem of language of scientific publications in the analyses. Although English is the most important language in written scientific communication, and can even be considered as the international language of science,[2] other languages are used as well. The *Science Citation Index* (SCI) contains relatively few non-English language journals from larger non-English speaking countries. One criticism of the outcomes presented by *May* addresses the problem of language-bias in the SCI by claiming that important non-English language journals are *not* covered by the SCI. Thus, indeed, this phenomenon may underrate the scientific output of larger countries with non-English domestic journals, such as Germany and France.

In this paper, however, we now present another, perhaps even more important dimension of the language-bias problem. Following some of the main lines adopted in an earlier paper,[3] we will show that the language-bias problem is also present *within* the boundaries of the SCI. Precisely this phenomenon is the main cause of the 'dramatically' modest impact figures of Germany and France.

Background and problem statement

The basic questions addressed in this paper are: how does the impact of papers written in other languages than English compare to that of articles written in English? And what is the effect of removing non-English language articles upon the value of impact indicators of institutions or entire countries? To which extent do rankings of countries according to their citation impact change when non-English papers are removed from the impact analysis? Which countries are most strongly affected?
These questions were addressed in three studies. The first relates to a scientific institute located in a non-English speaking country and active in the field of medicine.

The empirical material used is based on a bibliometric study for the Faculty of Medicine of the University of Münster in Germany. The time period analysed was 1990-1997. Data were drawn from the CD-Rom versions of the *Science Citation Index* (SCI) and the *Social Sciences Citation Index* (SSCI).

For the Faculty of Medicine in Münster, we calculated bibliometric indicators at several levels of aggregation, e.g., on the organizational level (faculty and institutes) and aggregated personnel level (for all scientific personnel aggregated and for the group of associate and full professors, having ranks C2-C4 in the German higher education system). Due to the large number of journals involved with German-language titles, we decided to analyze the data material according to publication language in the data set, namely English versus non-English.

In order to obtain a broader picture of the language bias problem, we have in a second study analyzed the output in medical research of five major Western countries: the USA, UK, France, Germany, and Switzerland. This part of our analysis focused on the role of language at a higher level of aggregation, namely that of national research systems, and more in particular the medical sciences. The data analyzed for the comparison of these five Western countries stems from the same data source. The period analyzed is much longer than that in the Münster study: 1981–1998.

Finally, in order to obtain an even broader overview, we expanded the analysis and added in a third study the output of Japan, Spain, and Italy to the output of Germany, France, Switzerland, included in the total *Science Citation Index*, thereby expanding the range of fields covered (now including in the analysis the natural and technical sciences as well), as well as the countries covered (now including in the analysis countries with comparable economic, demographic and scientific characteristics).

Methodology

In this section we discuss the variables and indicators used in our analysis. Some of the variables are directly derived from the ISI databases, particularly language, geographical location of the authors (country), document type and publication year. We define 'language' as the publication language indicated by ISI. 'Document type' relates to the document type mentioned in each publication in the ISI databases. In this study, we use only 'articles', 'letters', 'note's and 'reviews'. 'Country' is simply defined as the country appearing in the address heading of each publication within the ISI databases. A country gets the credits for a paper only once, even if there are more addresses from that country in the paper. 'Year' is defined as the year in which the paper was added to the database, or in other words, the database year. Finally, a 'year-block' consists of a number of subsequent 'years', thus setting the limits in the analysis for both publication and citation counts. We refer to earlier work for a detailed discussion of our bibliometric indicators.[4,5]

The impact indicator used in the study is denoted as CPP/FCSm. This indicator CPP/FCSm relates the measured impact Citation per Publication (CPP) of a specific entity (e.g., country or faculty) to a worldwide, field-specific reference value mean Field Citation Score (FCSm). In calculating FCSm, we used a the following procedure: FCSm is the mean citation rate of the fields in which the country has published, taking into account both the type of paper (e.g., normal article, review), as well as the specific years in which the papers were published. To give an example, the number of citations received during the period 1994–1998 by a letter published in 1994 in field X, is

compared to the average number of citations received during the same period (1994–1998) by all letters published in the same field (X) in the same year (1994). Of course, a country as a whole publishes its papers in many fields. Therefore, we calculated a weighted average indicated as FSCm, with the weights determined by the number of papers published in each field. Our definition of a 'field' is based on a classification of scientific journals into Journal Subject Categories developed by ISI. Although this classification is far from perfect, it provides a good first approximation, and is at present the only classification scheme directly compatible with the multidisciplinary structure of the data-system.

As mentioned above, we calculated indicators for overlapping year-blocks. For the Faculty of Medicine, University of Münster, we calculated indicators for four-year blocks (1990-1993, 1991-1994, … , until 1994–1997), while the medical research study for five Western countries was based on five-year blocks (1981–1985, 1982–1986, 1983–1987, … , until 1994–1998). This division in year-blocks generates a trend-analysis, with a moving average for CPP/FCSm. For example, in the medical research study of the five Western countries, the publications in the first year-block have variable citation windows, ranging from 5 years (for the 1981-publications) to 1 year (for the 1985-publications). This generates an average impact value for the whole set of publications for that year-block. The next year-block concerns the publications of 1982–1986. Now the publications of 1982 have a 5-year citation window, and the publications of 1983 have a 4-year citation window, and so on. This method allows for each publication to 'mature' in successive year-blocks. Together, such an analysis represents a trend-analysis up till as recent as possible (being 1998 at the moment of the study).

Results

First, the outcomes are presented of our analysis of the Faculty of Medicine at University of Münster (Germany). Figure 1 shows per year-block the total output of the faculty (including papers in all languages), and the output in English-language journal publications only. The share of the latter is about 80% of the former, which means that 20% of the output consists of non-English language journal publications

Figure 2 presents the citation impact scores related to the different types of output as presented in Figure 1. It becomes immediately clear that the impact related to *all* publications is considerably lower than the impact related to *only* the English language journal publications. This leads to the conclusion that the non-English-language

publications are lowering the impact of the faculty as a whole: the impact of the faculty increases to above worldwide average level due to the publications written in English, and particularly in the last two time-intervals of the analysis.

The next figures relate to the output in medical research of five major Western countries: USA, UK, France, Germany, and Switzerland. Figure 3 compares the medical research systems of the USA, UK, Germany, France and Switzerland based on the field-normalized impact indicator CPP/FCSm, in the period 1981–1998. The results of the analysis clearly indicate the modest positions of both Germany and France. Germany reaches an impact score slightly above worldwide average only in the most recent year-block.

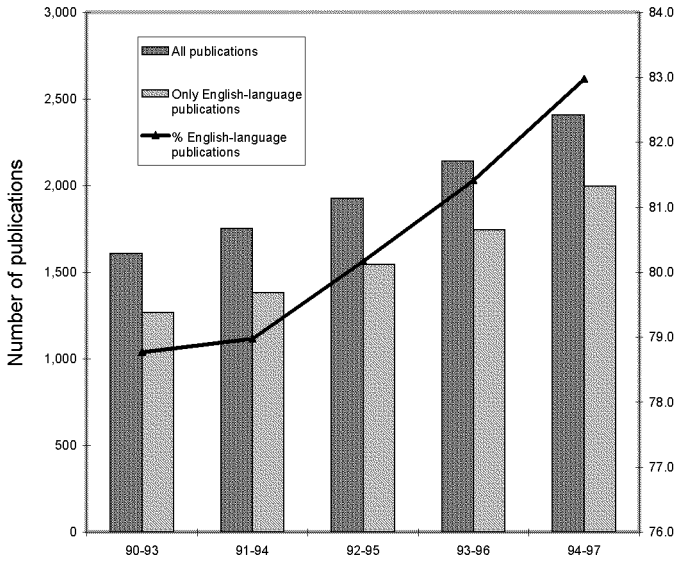

Figure 1. The output of the Faculty of Medicine, University of Münster (Germany), in all and in only English-language journal publications, 1990–1997

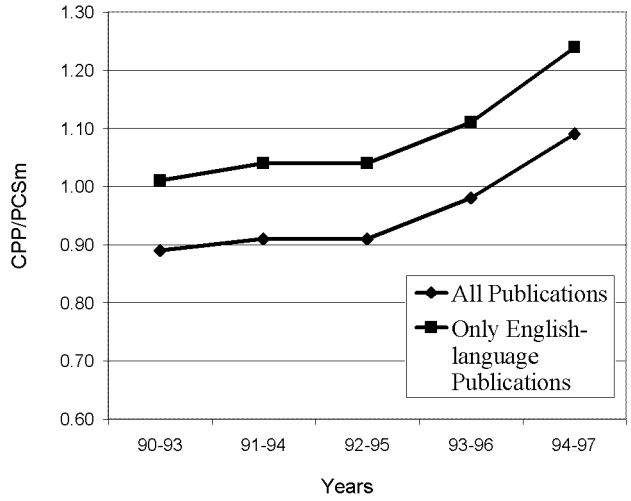

Figure 2. The impact scores of the Faculty of Medicine, University of Münster (Germany), in all and in only English-language journal publications, 1990–1997

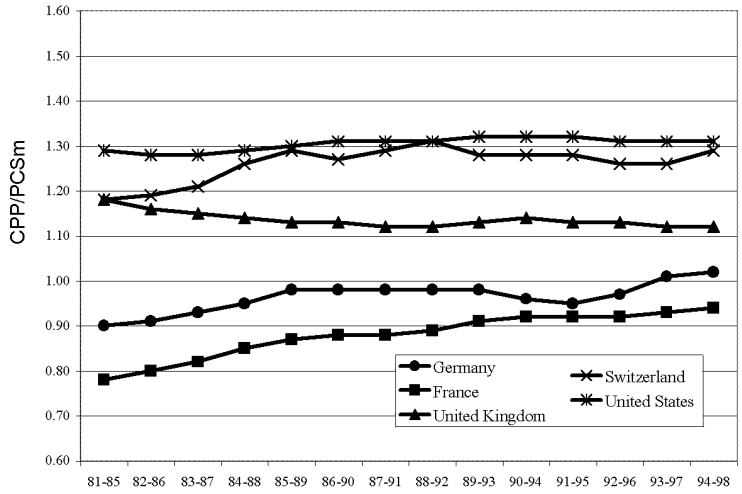

Figure 3. Impact analysis of five national medical R&D systems, all SCI-covered publications, medical research

In general terms our findings are in agreement with results published by *May*. Figure 4 presents the findings of the same analysis as in Figure 3, but now we eliminated the ISI-covered publications from (especially) German, French and Swiss medical scientists in other languages than English (though still 'covered' by the SCI). As is immediately visible in Figure 4, this restriction results in a very significant change: Switzerland now becomes the leading country in terms of impact scores, and Germany and France have impact scores above average, even competing with the UK.

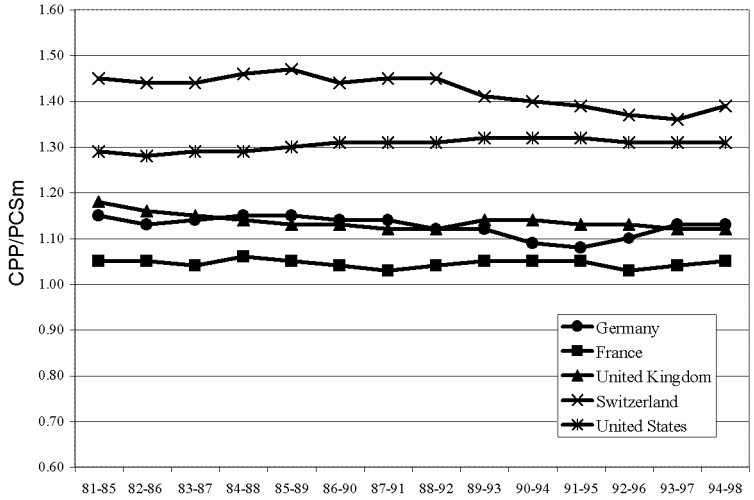

Figure 4. Impact analysis of five national medical R&D systems, only English language SCI-covered publications, medical research

To give an impression of the share of non-English language journals publications in relation to the total set of journal publications, Figure 5 displays the trend of this share for medical research for the USA, UK, Switzerland, Germany and France in the period 1981–1998. One clearly observes the diminishing share of non English-language output in the medical sciences for Switzerland, Germany and France, from about 30–40% to 10–15% in the period 1981–1998. However, 8% of the Swiss output in 1994–1998 is still published in non-English language journal publications, whereas Germany and France have shares around 15% in the same period. The relative share of non English-language journal publications of the USA and UK is close to zero, namely 0.1% of their output in the medical sciences.

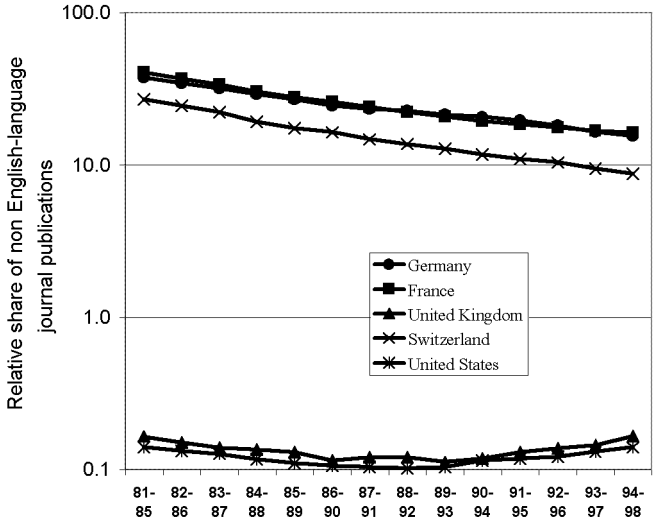

Figure 5. Relative share of non-English language journal publications in the total of all SCI-covered publications, medical research

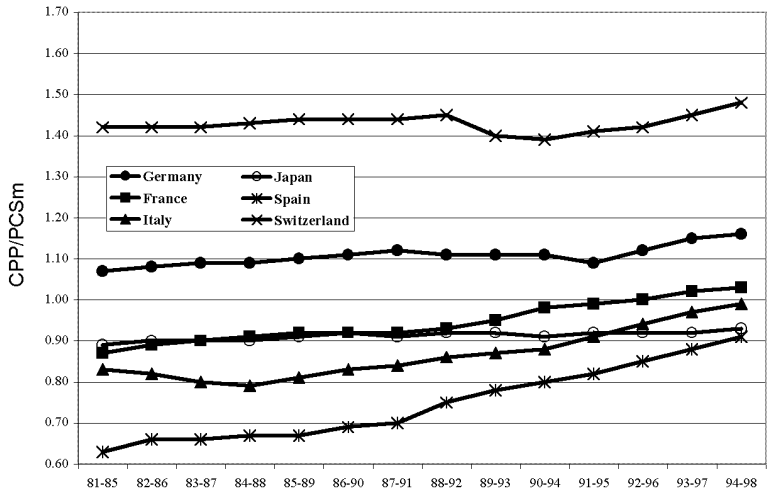

Figure 6. Impact analysis of six national research systems, all language SCI-covered publications, all fields of science

As a final test, we expanded the analysis and added the output of Japan, Spain, and Italy to the output of Germany, France, Switzerland, in the total *Science Citation Index*, thereby expanding the relevance of this methodological change both in terms of the fields covered (including in the analysis the natural and technical sciences, next to the medical sciences), as well as the countries covered (including in the analysis countries with comparable economic, demographic and scientific characteristics).

The results of the analysis, including the non-English language journal publications, are displayed in Figure 6. Switzerland has an impact of 1.4 in the period 1981–1998, while the German impact is about 1.1 (all fields of science). The other four countries have impact scores that differ in the earlier year-blocks of the analyzed period, but tend to converge in the latter periods of the analysis, with impact scores ranging between 0.9 and 1.0.

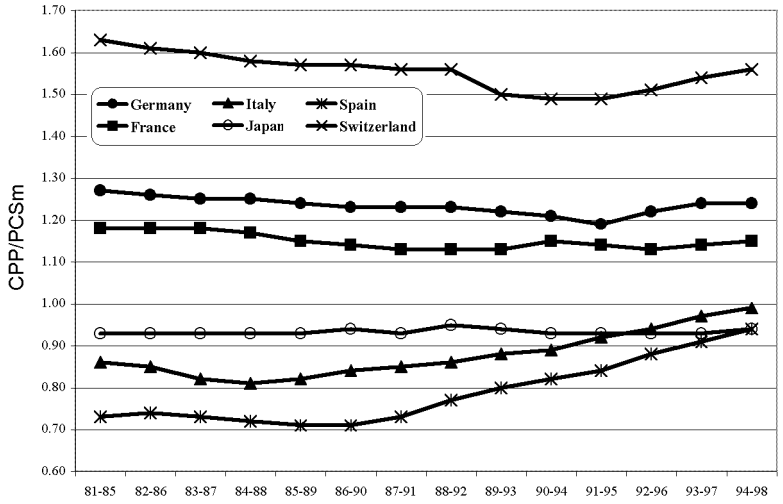

Figure 7. Impact analysis of six national research systems, only English-language SCI-covered publications, all fields of science

If one excludes the non-English language publications from the analysis we observe a similar effect as discussed for the medical sciences: impact scores of especially Switzerland, Germany and France increase significantly. These results are given in Figure 7. The effect is less for Italy, Spain and Japan. Apparently, the latter three

countries have fewer publications in other languages than in English within the scope of the *Science Citation Index*, whereas the former three countries, and especially France, have many publications in other languages than English.

Finally, we analyzed language coverage in the *Science Citation Index* as a whole over the period 1980–1998.[3] Figure 8 shows that the role of English as the major language of modern day science is striking. Publications in French, German, and particularly Russian show a decrease in the (relative) coverage in the *Science Citation Index* in the recent years, whereas the coverage of journal publications in other languages, grouped in 'Other', seems to have stabilized from 1993 onwards, with about 0.5 to 0.7 percent of all publications.

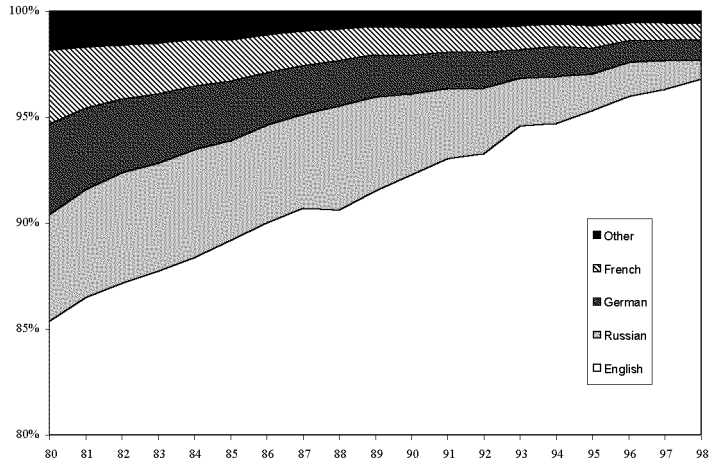

Figure 8. Language coverage within the *Science Citation Index*, 1980–1998

The changes in language coverage within the *Science Citation Index*, which becomes apparent from Figure 8, can be partly explained by a number of possible explanations, like for example by focusing on the national orientation of scientific journals within the database and shifts in this orientation, or on a decreasing share of covered scientific journals from non-English speaking countries in the database. However, the influences of these factors remains to be seen, and surely exceeds the boundaries of the current investigation. Forthcoming analyses will throw light on this phenomenon.

Concluding remarks

We have shown empirically that language biases play an important role in the comparison and evaluation of national science systems, due to the simple fact that the SCI also covers non-English language journals of which the papers have a considerably lower impact than those in the English-language journals. Additional analyses have shown that the US and UK have barely any publications in other than English-language journals, whereas countries such as Germany, France and Switzerland have a decreasing though still significant number of publications in other languages than English. We therefore conclude that this language aspect within the boundaries of the SCI seriously (but 'artificially') 'dilutes' the impact score of these major scientific nations. The impact scores of Germany, France and Switzerland show increases above at least 10%, but larger increases were observed as well.

However, the exclusion of the non-English journal publications will, in combination with for example demographic data, decrease both the number of publications per capita as well as the total impact per capita (i.e., not field-normalized). This study thus clearly indicates the sensitivity of bibliometric data, and the importance of careful analyses when combining this type of material with other quantitative data.

A more detailed analysis should address whether or not the publication language is solely responsible for the observed lower impact of non-English language publications. Perhaps German, French and Swiss scientists tend to publish specific research communications towards their national professional colleagues in domestic language journals. In addition, the influence on the impact scores of publications by English and US scientists in journals with a supposed strong national orientation should be investigated as well, and probably left out of impact analyses as well.

Furthermore, it should also be noted that in the parts of the analysis in which we removed the non-English language publications, we removed these publications only from the publication output. *Citations in* these no-English language publications to articles in English were *not* deleted. The effect of removing also non-English language publications as *sources of citations* should also be addresses in a follow-up study.

These considerations clearly illustrate that bibliometric indicators even at the macro-level should not be taken at face value and need to be interpreted against the background of their inherent limitations, such as the language of research papers.

References

1. R. M. MAY, The scientific wealth of nations, *Science*, 275 (Feb 7) (1997) 793–796.
2. E. GARFIELD, English – An international language for science, *The Information Scientist*, Dec 76, In: *Essays of an Information Scientist*, Vol. 1, pp. 19–20, ISI Press.
3. T. N. VAN LEEUWEN, H. F. MOED, R. J. W. TIJSSEN, M. S. VISSER, A. F. J. VAN RAAN, First Evidence of serious language-bias in the use of citation analysis for the evaluation of national science systems, *Research Evaluation*, Vol. 9, August 2000, 155–156.
4. H. F. MOED, R. E. DE BRUIN, Th. N. VAN LEEUWEN, New bibliometric tools for the assessment of national research performance: database description, overview of indicators and first applications, *Scientometrics*, 33 (1995) 381–422.
5. A. F. J. VAN RAAN, Advanced bibliometric methods as quantitative core of peer review based evaluation and foresight exercises, *Scientometrics*, 36 (1996) 397–420.

Received February 16, 2001.

Address for correspondence:
THED N. VAN LEEUWEN
Centre for Science & Technology Studies (CWTS),
Leiden University, Leiden, The Netherlands
E-mail: leeuwen@cwts.leidenuniv.nl