Jointly published by Akadémiai Kiadó, Budapest
and Springer, Dordrecht

Scientometrics,
Vol. 66, No. 1 (2006) 71–80

The influence of author self-citations on bibliometric meso-indicators. The case of european universities

BART THIJS,[a] WOLFGANG GLÄNZEL[a,b]

[a] *Steunpunt O&O Statistieken, Katholieke Universiteit Leuven, Leuven (Belgium)*
[b] *Institute for Research Organization, Hungarian Academy of Sciences, Budapest (Hungary)*

In earlier studies by the authors, basic regularities of author self-citations have been analysed. These regularities are related to the ageing, to the relation between self-citations and foreign citations, to the interdependence of self-citations with other bibliometric indicators and to the influence of co-authorship on self-citation behaviour. Although both national and subject specific peculiarities influence the share of self-citations at the macro level, the authors came to the conclusion that – at this level of aggregation – there is practically no need for excluding self-citations.

The aim of the present study is to answer the question in how far the influence of author self-citations on bibliometric meso-indicators deviates from that at the macro level, and to what extent national reference standards can be used in bibliometric meso analyses. In order to study the situation at the institutional level, a selection of twelve European universities representing different countries and different research profiles have been made. The results show a quite complex situation at the meso-level, therefore we suggest the usage of both indicators, including and excluding self-citations.

Introduction

In the ongoing debate about 'correct' citation counting the issue of author self-citations has often been raised. Policy makers and researchers question at what level of aggregation self-citations may cause distortions and affect the validity of bibliometric indicators (see AKSNES, 2003).

In earlier macro studies, basic regularities of author self-citations have been studied by the authors (GLÄNZEL et al., 2004; GLÄNZEL & THIJS, 2004). The authors have found that the share of self-citations considerable differs among the individual countries. Although subject specific characteristics and national peculiarities could be found, the authors came to the conclusion that at the macro level there is practically no need for eliminating self-citations in evaluative bibliometrics.

Received June 22, 2005

Address for correspondence:
BART THIJS
Katholieke Universiteit Leuven, Steunpunt O&O Statistieken, Dekenstraat 2, B-3000 Leuven, Belgium
E-mail: Bart.Thijs@econ.kuleuven.ac.be

0138–9130/US $ 20.00

Observations and experience made at the micro level, particularly, in the context of the publication output of individuals and research team/departments, however, show that this statement might not hold at lower levels of aggregation. CWTS is usually excluding self-citations because their inclusion might form an important source of error in the ratio of observed/expected citation impact (NEDERHOF et al., 1993). CWTS uses self-citation rates to detect departments with divergent levels of self-citation.

The aim of this study is to answer the question in how far the influence of author self-citations on bibliometric meso-indicators deviates from that at the macro level, and to what extent national reference standards can be used in bibliometric meso analyses.

Data sources and processing

The results of this study are based on raw bibliographic data extracted from the 2000-2001 annual cumulations of the Web of Science (WoS) of the Institute for Scientific Information (ISI – Thomson Scientific, Philadelphia, PA, USA). The extracted data have undergone a detailed cleaning and then processed to bibliometric indicators. All papers of the document type *articles, letters, notes* and *reviews* have been taken into consideration. Citations received by these papers have been determined for three-year citation window beginning with the publication year on the basis of an item-by-item procedure using special identification-keys (so-called cluster-keys) made up of bibliographic data elements. Citation data up to 2003 has been extracted from the WoS.

For the present study, the same definition of self-citations has been applied as was used in earlier studies by the authors and which is basically identical to that introduced by SNYDER & BONZI (1998) and applied by AKSNES (2003). According to this definition, a self-citation occurs whenever the set of co-authors of the citing paper and that of the cited one are not disjoint, that is, if these sets share at least one author. We would like to stress again that the reliability of this methodology is affected by homonyms and spelling variances/misspellings of author names.

Subject classification of the publications was based on the field assignment of journals according to sixteen major fields of science developed in Leuven and Budapest (GLÄNZEL & SCHUBERT, 2003). These fields are Agriculture & Environment, Biology (Organismic & Supraorganismic Level), Biosciences (General, Cellular & Subcellular Biology, Genetics), Biomedical Research, Clinical and Experimental Medicine I (General & Internal Medicine), Clinical and Experimental Medicine II (Non-Internal Medicine Specialties), Neuroscience & Behaviour, Chemistry, Physics, Geosciences & Space Sciences, Engineering, Mathematics, Social Sciences I (General, Regional & Community Issues), Social Sciences II (Economical & Political Issues) and Arts & Humanities. Journals can also be classified as Multidisciplinary.

Papers were assigned to countries and institutes based on corporate addresses given in the by-line of the publication. For all European countries citation-based macro-indicators have been calculated, where two different sets of indicators have been build, namely, using self-citations and excluding them. Six medium-sized European countries representing regions from northern, central and western Europe have been selected: Austria, Denmark, Finland, Hungary, Ireland and Sweden. All institute names as they appear in WOS of the selected countries have been cleaned-up on the basis of both computerised and manual procedures. A thesaurus has been made up of cleaned names with all theirs spelling variances. The correct institutional assignment of addresses has been checked by national experts. The only formal requirement for the selection of institutes for the meso-analysis was a minimum publication output of 10 publications over the 2-year period 2000/2001.

Methods

This study consists of two parts. First, citation indicators on a macro-level were calculated. In particular, the following set of indicators was used:

- MOCR, Mean Observed Citation Rate. This indicator is defined as the ratio of citation count to publication count. MOCR reflects the factual citation impact of unit (here: country or institution) including self-citations.
- MOCRX, Mean Observed Citation Rate eXcluding self-citations. If we denote the share of self-citations of a unit (country or institution) by S, we obtain MOCRX = (1-S)·MOCR from the definition.
- MECR, Mean Expected Citation Rate, is a journal-based indicator expressing an expected citation rate of a given paper set. The (journal-based) expected citation rate of a single paper is defined as the average citation rate of all papers published in the same journal in the same year. Instead of the one-year citation window to publications of the two preceding years as used in the *Journal Citation Report* (JCR), a three-year citation window to one source year is used, as explained above. For a set of papers assigned to a given country, region or institution in a given field or subfield, the indicator is the average of the individual expected citation rates over the whole set.
- MECRX, Mean Expected Citation Rate eXcluding self-citations, is a special journal impact measure expressing the expected rate of citations excluding the self-citations. It's calculated similar to the MECR and has the same implications as the previous one.
- RCR, Relative Citation Rate, is the ratio of MOCR and MECR. RCR thus compares the observed citation rate with the expected one. RCR = 0

corresponds to uncitedness, RCR < 1 means lower-than-average, RCR > 1 higher-than-average citation rate, RCR = 1 if the set of papers in question attracts just the number of citations expected on the basis of the average citation rate of the publishing journals

- RCRX, Relative Citation Rate eXcluding self-citations, is build analogously to the previous indicator, particularly, RCRX = MOCRX/MECRX.

Indicators are calculated on aggregation within countries and within subject fields across the selected countries.

In order to build valid measures, all indicators are based on the same publication period (2000–2001) and the same citation window (3 year, year of publication plus 2 successive years). Only the RCR and RCRX will be reported in this paper.

In the second part of this study indicators for twelve selected institutes were calculated.

First publications had to be assigned to institutes and research profiles for each organization had to be calculated. Four steps were taken to complete this task.

1. *Address cleaning* (as described in the previous section).
2. *Address assignment.* The thesaurus obtained in the previous step was matched with the address data in WOS.
3. *Research profile.* For each institute, the percentage of publications within each of the sixteen subject fields listed in the previous section was calculated.
4. *Selection of most productive institutes.* Only institutes with at least 10 publications in the period (2000–2001) are selected.

After these steps, factor analysis and clustering algorithms have been applied. The objective was to create groups of institutions with similar research profiles in order to compare likewise institutions. The percentages of publications in each research field are used as input data. This data is standardized (0–1). Therefore, the total number of publications of an institute has no effect on the final clustering.

This standardized data resulted, through PCA with varimax rotation, in 9 components. Component scores were calculated and used for hierarchical clustering (squared Euclidean distance with Ward linkage).

Institutions within the same group thus represent similar profiles; different groups represent rather different research and publication profiles. This selection allows both, the cross-national comparison of institutions with similar profiles and national comparison of institutions with different profiles on the basis of the share of self-citations and other citation indicators in all fields combined. The analysis in major fields allows cross-group comparisons, too.

In all, twelve institutions representing different clusters and countries have been chosen for further analysis. The institutions will, however, be treated anonymously; names are substituted by numbers.

Results

The first part of the analysis yielded results at the national level. Table 1 presents the share of self-citations as well as the two relative citation rates (including and excluding self-citations) for the six selected countries. The indicator values are in line with those determined for 1999 (cf. GLÄNZEL et al., 2004; GLÄNZEL & THIJS, 2004) although the relative citation rates of some countries have slightly increased. Four of the six countries represent cases where observed citation impact exceeds expectation, one country has an observed citation rate that meets the expected and one where impact remains below its expectation.

Table 1. National citation-based indicators for six selected European countries in all fields combined (Publication period: 2000–2001; 3-year citation window)

Country	Share of self-citations	RCR	RCRX
Austria	29%	1.075	1.025
Denmark	28%	1.196	1.163
Finland	29%	1.151	1.105
Hungary	35%	0.937	0.868
Ireland	25%	1.217	1.248
Sweden	27%	1.138	1.116

At a first sight one could conclude that the relative citation rate might always decrease if self-citations are removed, but this is certainly not the case since for the world total we have by definition RCR = RCRX = 1. Ireland is, for instance, one of the counties the relative citation rate of which grows if self-citations are removed.

Table 2 presents the same indicators for the sixteen subject fields over the six selected countries. All these fields, except ‘Arts and humanities’, have a relative citation rate (including or excluding self-citations) that is higher than the world average, which is 1 for all fields. However this is completely in line with the higher values in Table 1. The differences in the share of self-citations between the fields has been reported by the authors (cf. GLÄNZEL & THIJS, 2004). There’s no clear effect of the share of self-citations on the impact of the exclusion of self-citations from the calculation of the relative citation rate. Agriculture and Engineering, for instance, have the same share (30%) of self-citations, however, the exclusion has an opposite effect on the RCR.

Table 2. Citation-based indicators for sixteen subject fields in the six selected countries (Publication period: 2000–2001; 3-year citation window)

Field	Share of self-citations	RCR	RCRX
Agriculture	30%	1.178	1.149
Arts & humanities	25%	0.950	0.932
Biology	26%	1.065	1.032
Biomedical research	23%	1.050	1.016
Biosciences	20%	1.061	1.037
Chemistry	31%	1.055	1.036
Engineering	30%	1.196	1.216
General & internal medicine	16%	1.152	1.139
Geoscience	30%	1.124	1.074
Mathematics	38%	1.075	1.081
Multidisciplinary	12%	1.192	1.176
Neuroscience and behaviour	22%	1.044	1.020
Non-internal med. specialties	20%	1.155	1.121
Physics	32%	1.087	1.065
Social sciences I	25%	1.095	1.051
Social sciences II	20%	1.056	1.044

In the second part of the analysis, the share of publications within each of the sixteen fields is calculated for all individual institutions in the selected countries. These scores were used in a PCA resulting in 9 components. These components accounted for 76,7% of the total variance in the data. After rotation of these components, scores were calculated for all institutes.

Hierarchical clustering with squared Euclidean distances and Ward-linkage was used to create clusters of likewise institutes. The applied method resulted in six different clusters, particularly, in clusters with natural sciences, biomedical research, social sciences and humanities, clinical and experimental research, bio–environmental research and specialised medical specialties, respectively, in the main focus.

Table 3 presents the share of publications within each subject field for each group. Dominant fields are highlighted. On the basis of Table 3 we could characterise the profile of the six clusters as shown below where we also indicate the number of institutes in each cluster.

Table 3. Share of publications within each subject field

Field	Cluster A	Cluster B	Cluster C	Cluster D	Cluster E	Cluster F
Agriculture	5%	**38%**	2%	5%	0%	9%
Arts & humanities	0%	0%	0%	3%	0%	0%
Biology	4%	**48%**	5%	5%	0%	2%
Biomedical research	2%	2%	11%	3%	5%	0%
Biosciences	4%	11%	12%	3%	3%	0%
Chemistry	**34%**	7%	5%	5%	0%	1%
Engineering	**25%**	1%	1%	13%	0%	4%
General & internal medicine	2%	3%	**30%**	1%	6%	0%
Geoscience	3%	6%	1%	1%	0%	**81%**
Mathematics	6%	0%	0%	6%	0%	0%
Multidisciplinary	0%	0%	0%	0%	0%	0%
Neuroscience and behaviour	1%	0%	3%	3%	**48%**	0%
Non-internal med. specialties	6%	4%	**41%**	2%	**63%**	0%
Physics	**21%**	0%	2%	1%	0%	3%
Social sciences I	1%	0%	2%	**29%**	0%	0%
Social sciences II	2%	0%	0%	**42%**	0%	2%

Cluster A: Eng/Phys/Chem –Technical research institutes: Engineering, Physics and Chemistry (105 institutes)
Cluster B: Agri/Bio – Agricultural institutes (55 institutes)
Cluster C: GenMed – Research institutes with main focus on medical research (221 institutes)
Cluster D: NonMed – Institutes with a multidisciplinary profile, without medical research (17 institutes)
Cluster E: SpecMed – Institutes for specialized medicine (13 institutes)
Cluster F: Geo – Geosciences and Space sciences (17 institutes)

Table 4. Citation-based indicators for six clusters obtained in the second step
(Publication period: 2000–2001; 3-year citation window)

Cluster	Share of self-citations	RCR	RCRX
A	35%	1.045	1.007
B	35%	1.084	1.014
C	27%	1.119	1.084
D	29%	1.068	1.050
E	26%	1.215	1.171
F	39%	0.963	0.871

Table 4 presents the citation-based indicators for the six clusters. The share of self-citations is in keeping by and large with the macro data published in GLÄNZEL & THIJS (2003), although in cluster E and F the share in the selected countries is higher that expected. Both the RCR and RCRX values in clusters A, B and D correspond to the world average; RCR in cluster F is also close to the standard, while its RCRX value is rather low. Here author self-citation seems to distort the picture a bit.

On the other hand, both RCR and RCRX values in cluster E are above all standards. Thus, some of the selected countries seem to be strong in miscellaneous medical specialities.

For the calculation of the impact of author self-citations on the meso-level, twelve research institutes have been selected. These criteria were taken into account by selecting the institutes:

– a reasonable number of publications in the 2-year period,
– evenly distributed across the clusters (Due to the number of members of cluster C, 3 institutes were selected from this group and only 1 was selected from cluster E),
– evenly distributed across countries.

Table 5 shows the selection, and visualises the assignment of the institutes to countries and profile clusters.

Table 5. Location and profile of the selected institutes

Cluster	Austria	Denmark	Finland	Hungary	Ireland	Sweden
A			6	8		
B		3			9	
C	1				10	12
D	2					11
E				7		
F		4	5			

Table 6 shows that several institutes have a relatively high share of author self-citations; moreover, the difference between RCR and RCRX also shows that excluding the author self-citations results in considerable changes of the indicator values.

Table 6. Citation-based indicators for the twelve selected institutes

Institute	Country	Cluster	Share of self-citations	RCR	RCRX
1	A	GenMed	30%	1.064	0.984
2	A	NonMed	17%	1.124	1.341
3	DK	Agri/Bio	38%	1.069	1.015
4	DK	Geo	40%	1.091	0.981
5	FIN	Geo	49%	1.022	0.811
6	FIN	Eng/Phys/Chem	41%	1.092	0.970
7	H	SpecMed	19%	1.135	1.176
8	H	Eng/Phys/Chem	41%	0.930	0.827
9	IRE	Agri/Bio	37%	1.308	1.232
10	IRE	GenMed	19%	1.666	1.792
11	S	NonMed	29%	1.281	1.156
12	S	GenMed	27%	1.112	1.069

Some of the selected institutions clearly deviate from both the national standard and the "cluster standard" in the country group (cf. Table 6). Several institutes are on one hand characterised by very low shares of self-citations and, on the other hand, by a strong increase of relative citation impact if self-citations are removed. In this context, the high relative citation rates with low shares of self-citations in institutes 2 (Austria), 7 (Hungary) and above all 10 (Ireland) with RCRX = 1.79 are worth mentioning. Relative low share of self-citation is, however, no guarantee for improving citation impact by excluding self-citations, as the example of institutes 11 and 12 in Sweden shows. Institutes 5 (Finland) and 8 (Hungary) show the opposite direction of the above-mentioned trend. Others, for instance, institute 1 in Austria and 12 in Sweden are truely representative for their country and publication profile. And a relatively high share of self-citations does not necessarily go with lower RCRX value. Institute 9 in Ireland may serve just as an example. Although the S value is higher than the Irish average, the RCRX value is very close to the national standard.

The above example shows that at the institutional level the additional analysis of author-self citations might serve as a valuable and indispensable tool to providing additional information to evaluative citation studies.

Conclusions

The results of the institutional analysis partially deviate both, from each other, from their field standard and the corresponding national standard. This reflects a quite complex situation: Research at the meso level is, on one hand, more characterised by specific profiles than the national level is. On the other hand, institutional research is less specialised than research in smaller units such as departments, teams or even that of individuals, and thus less affected by topic characteristics or by the communication behaviour of the most profilic authors representing the group. The main conclusion of this study is in line with the recommendation by AKSNES (2003) who argued that at lower levels of aggregations, such as the meso-level, self-citations might represent a serious problem. He recommends to preferably removing self-citations before making comparisons or, at least, to carefully consider by-effects caused by self-citations before using citations as indicators of scientific impact. Given the utmost interesting macro and meso figures presented in Tables 1, 3 and above all, in Table 5 we would like to conclude that at the meso level, the share of self-citation, the corresponding citation-based indicators both including and excluding self-citations should be presented along with their national and subject standards to better understand the complexity of scientific communication as reflected citation impact of universities and research institutes.

The results presented here are, of course, of preliminary nature as they are restricted to a selection of European institutions. The extension of this analysis to a broader set of research institutions is the task of later studies.

References

AKSNES, D.W. (2003), A macro-study of self-citations, *Scientometrics*, 56 (2) : 235–246.

GLÄNZEL, W., SCHUBERT, A. (2003), A new classification scheme of science fields and subfields designed for bibliometric evaluation purposes, *Scientometrics*, 56 (3) : 357–367.

GLÄNZEL, W., THIJS, B., SCHLEMMER, B. (2004), A bibliometric approach to the role of author self-citations in scientific communication, *Scientometrics*, 59 (1) : 63–77.

GLÄNZEL, W., THIJS, B. (2004), World flash on basic research – the influence of author self-citations on bibliometric macro indicators, *Scientometrics*, 59 (3) : 281–310.

NEDERHOF, A. J., MEIJER, R. F., MOED, H. F., VAN RAAN, A. F. J. (1993), Research performance indicators for university departments - A study of an agricultural university, *Scientometrics*, 27 (2) : 157–178.

SHARP, D. (2004), As we said ..., *Lancet*, 364 (9436) : 744–744.

SNYDER, H., BONZI, S. (1998), Patterns of self-citation across disciplines, *Journal of Information Science*, 24 : 431–435.