The Weight of Author Self-citations

Wolfgang Glänzel[*,**], Bart Thijs[*] and András Schubert[**]

[*]*Wolfgang.Glanzel@econ.kuleuven.ac.be, Bart.Thijs@econ.kuleuven.ac.be*
Steunpunt O&O Statistieken, KU Leuven, Dekenstraat 2, B-3000 Leuven (Belgium)

[**]*glanzw@helka.iif.hu, schuba@helka.iif.hu*
Institute for Research Policy Studies, Hungarian Academy of Sciences, P.O.B. 994, H-1245 Budapest (Hungary)

Abstract
The discussion about how to treat author self-citations driven by policy application and quality measurement intensified in the last years. The definition introduced by Snyder and Bonzi has – in lack of any reasonable alternative – been used in bibliometric practice for science policy purposes. This method, however, does not take into account the weight of self-citing authors among co-authors of both the cited and citing papers. The objective of the present paper is to quantify the weight of self-citations with respect to co-authorship.
The analysis is conducted at two levels: at the macro level, namely, for fifteen subject fields and the most active forty countries, and at the meso level, for a set of selected research institutions.

Introduction

The ongoing debate on interpretation, role and treatment of author self-citations in bibliometrics and policy use has been intensified. Science policy regards the citation as part of a reward system; self-citations consequently distort the system as such. Information science interprets citations as part of communication in science. The debate has thus resulted in a certain polarisation.

Recent bibliometric studies have aimed at analysing the role of author self-citation as viewed from the perspective of bibliometric methodology, namely, at applying a quantitative, statistical approach. From this perspective, regularities related to the ageing of self-citations, as well as to their relation with foreign citations and with other bibliometric indicators have been found which allow the conclusion that at the macro level self-citations may be considered a natural part of scientific communication, indeed. However, meso-studies (e.g., *Aksnes* 2004, *Nederhof* et el., 1993) have shown that inclusion of self-citations might form a source of error, for instance, in the ratio of observed/expected citation impact. The Centre for Science and Technology Studies (CWTS) at Leiden University (The Netherlands), for instance, uses self-citation rates to detect departments with deviant levels of self-citation. Although information scientists and bibliometricians have – as shown by these examples – paved the way for a pragmatic discussion, the policy-driven approach to author-self citations as used in research evaluation and calculation of funding formulas results in the interpretation as a source of distortion of the impact of scientific research. Nevertheless, results of bibliometric studies are increasingly used in policy-relevant context; bibliometricians must therefore take the responsibility for these applications, too. In the context of author self-citations this means also to examine the possibility that the relatively rough measure as defined by Snyder and Bonzi might be refined to provide a measure compensating for the uneven weight of author self-citations caused by unequal co-authorship patterns.

The present study provides a large-scale analysis of the share and the ageing of self-citations, as well as a breakdown by science fields on the basis of the total publication output indexed in selected annual updates of the *Web of Science*®. In a second step, the proposed method will be applied to a selection of universities and research institutions of different research profiles.

Data sources and data processing

The results of this study are based on raw bibliographic data extracted from the 1994-2003 annual cumulations of the *Web of Science*® (WoS) of the *Institute for Scientific Information* (ISI – Thomson Scientific, Philadelphia, PA, USA). The extracted data have carefully been cleaned and then processed to bibliographic indicators. All papers of the document type *articles, letters, notes* and *reviews* indexed in the 1994 and 2000 annual updates of the WoS have been taken into consideration. Citations

Proceedings of ISSI 2005, pages 382–389.
Edited by P. Ingwersen and B. Larsen.

received by these papers have been determined for the period beginning with the publication year till 2003 on the basis of an item-by-item procedure using special identification-keys made up of bibliographic data elements. Papers were assigned to countries based on the corporate address given in the by-line of the publication. All countries indicated in the address field were thus taken into account. Subject classification of publications was based on the field assignment of journals (in which the publications in question appeared) according to the twelve major fields of science and three fields of social sciences and humanities developed in Leuven and Budapest (see, for instance, *Glänzel* and *Schubert*, 2003). In particular, the following fields have been used: 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 & Behavior, Chemistry, Physics, Geosciences & Space Sciences, Engineering, Mathematics and Social Sciences I (General, Regional & Community Issues), Social Sciences II (Economical & Political Issues) and Arts & Humanities, respectively.

Methods and results

In bibliometric studies, the definition of self-citations by *Snyder* and *Bonzi* (1998), *Aksnes* (2002) and *Glänzel* et al. (2004) is applied. 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. Although, the reliability of this method is affected by homonyms (resulting in Type II errors by erroneous self-citation counting) and spelling variances/misspellings of author names (resulting in Type I errors by not recognising self-citation), – at the meso and macro level – there is no feasible alternative. Even if we might assume that the two types of errors balance out at higher levels of aggregation, the question arises of what the real weight of a self-citation is. The weight of a self-citation might be influenced by the contribution of the self-citing co-author(s) in the total of all co-authors involved. The self-citation link between the citing and cited work is obviously much stronger if a single-authored paper is cited in a single-authored paper of the same author than if this link is, for instance, created between two multi-authored papers by only one joint co-author. The question arises of the binary self-citation count, namely, '1' if a self-citation occurred and '0' if this is not the case, should be replaced by a continuous measure reflecting a 'fuzzy' situation.

Figure 1 visualises four different situations of author self-citations with not-empty sets of co-authors ($\mathbf{A} \neq \emptyset \wedge \mathbf{B} \neq \emptyset$). Case a) corresponds to a foreign citation, that, is no self-citation ($\mathbf{A} \cap \mathbf{B} = \emptyset$), b) complete self-citation, that is, all co-authors of the citing paper are also the co-authors of the cited work ($\mathbf{A} = \mathbf{B}$**),** c) 'partial' self-citation, for instance, all co-authors of the citing paper are among the co-authors of the cited work or the opposite case, but the two sets do not coincide ($\mathbf{A} \subset \mathbf{B}$ or $\mathbf{A} \supset \mathbf{B}$**)** and d) citing and cited work share some but not all co-authors ($\mathbf{A} \backslash \mathbf{B} \neq \emptyset \wedge \mathbf{B} \backslash \mathbf{A} \neq \emptyset$).

A potential measure should take both, the number of co-authors of cited and citing work and the number of joint co-authors in these sets, into account. *McNee et al.* (2002) used a cosine similarity metric in the context of *Collaborative Filtering* in the recommending of citations for research papers. Indeed, Salton's cosine measure ($r_{\mathbf{AB}}$) is a possible measure of relative self-citation if network properties are studied. However, if self-citation links between two individual papers are analysed, the most appropriate candidate is beyond any doubt the Jaccard Index ($J_{\mathbf{AB}}$). $J_{\mathbf{AB}}$ is the ratio of the cardinality of the intersection of two sets **A** and **B** and the cardinality of their union, particularly $J_{\mathbf{AB}} = |\mathbf{A} \cap \mathbf{B}|/|\mathbf{A} \cup \mathbf{B}|$. In verbal terms, the Jaccard Index relates the number or 'weight' of co-authors contributing to both, cited and citing papers to the number of all co-authors of the two publications.
In general, we know that $J_{\mathbf{AB}} \leq r_{\mathbf{AB}}$ with $J_{\mathbf{AB}} = r_{\mathbf{AB}}$ *iff* $\mathbf{A} = \mathbf{B}$ or $\mathbf{A} \cap \mathbf{B} = \emptyset$ (**A** and **B** are not empty). In the four examples in Figure 1 we have: *a)* $J_{\mathbf{AB}} = r_{\mathbf{AB}} = 0$, *b)* $J_{\mathbf{AB}} = r_{AB} = 1$, *c)* $J_{\mathbf{AB}} = 0.750$; $r_{\mathbf{AB}} = 0.866$ and *d)* $J_{\mathbf{AB}} = 0.167$; $r_{\mathbf{AB}} = 0.289$. We will call such self-citation measures *fractional* self-citation counting in contrast to the traditional *binary* counting proposed by Snyder and Bonzi.

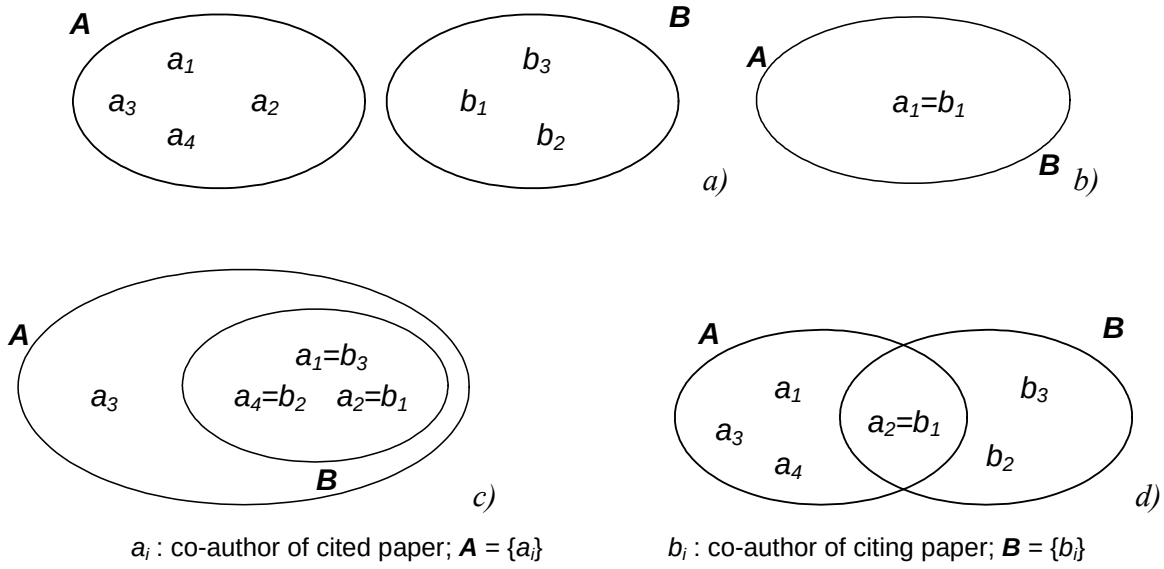

Figure 1 Four typical situations of author self-citation
(a) 'foreign' citation, b) 'complete' self-citation, c) and d) cases of 'partial' self-citation)

Fractional self-citation counting based on the Jaccard Index will be applied to a 10-year prospective (diachronous) citation analysis of the 1994 volume of the *Web of Science*®. Figure 2 presents the life-time distributions of self-citations (both integer and fractional count) and all citations based on the 10-year period for all fields combined. The life-time distributions are calculated on the basis of empirical values of increments with respect to the total of corresponding self-citations and citations received during the ten years. The ageing of binary self-citations is obviously much faster than that of all citations. Ageing is even faster if a fractional counting scheme is applied. The peak is reached in the third year after publication; from the fourth year on fractional self-citation rates drastically decrease.

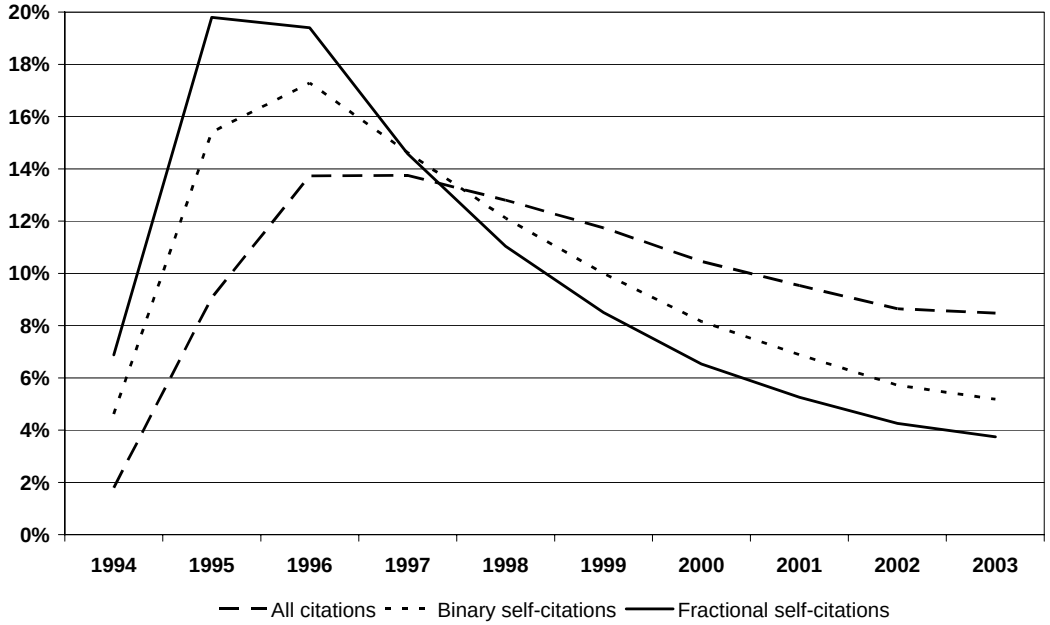

Figure 2 Empirical density of the life-time distribution of different forms of citations for all science fields combined

Obviously, the share of self-citations decreases if fractional counting is applied. Figure 2 presents the breakdown of binary and fractional self-citation count (the latter one based on the Jaccard Index) by 15 subject fields in the sciences, social sciences and humanities. The key to field abbreviations and

subject codes can be found in the Appendix. As expected, the relationship between share of fractional and binary self-citations is dependent of the field since collaboration characteristics widely differ among the fields. In order to visualise this effect, a variant of the *Collaborative Coefficient* (*CC*) according to *Ajiferuke* et al. (1988) has been used. $CC^* = 1 - CC$, being a harmonic mean, expresses the co-authors' average contribution in their papers in a given set of publications. $CC^* = 1$ *iff* all papers are single-authored. By contrast, CC^* takes small positive values around 0, for instance, in high-energy physics with traditionally very high number of co-authors. However, large extent of collaboration, i.e., low CC^* values do not automatically imply low share of self-citations as well. This is in line with those results by *Glänzel* and *Thijs* (2004), namely, that the number of co-authors does not inflate the share of self-citations. This observation applies also to the fractional case. Physics and Chemistry with low CC^* values and high share of binary/fractional self-citations, on one hand, and Social Sciences and Humanities with low collaboration and low binary/fractional self-citations, on the other hand, might just serve as examples. In Biosciences and Geosciences the share of self-citation drastically decreases if the fractional scheme is applied; in Mathematics and Social Sciences II, the effect is much less pronounced.

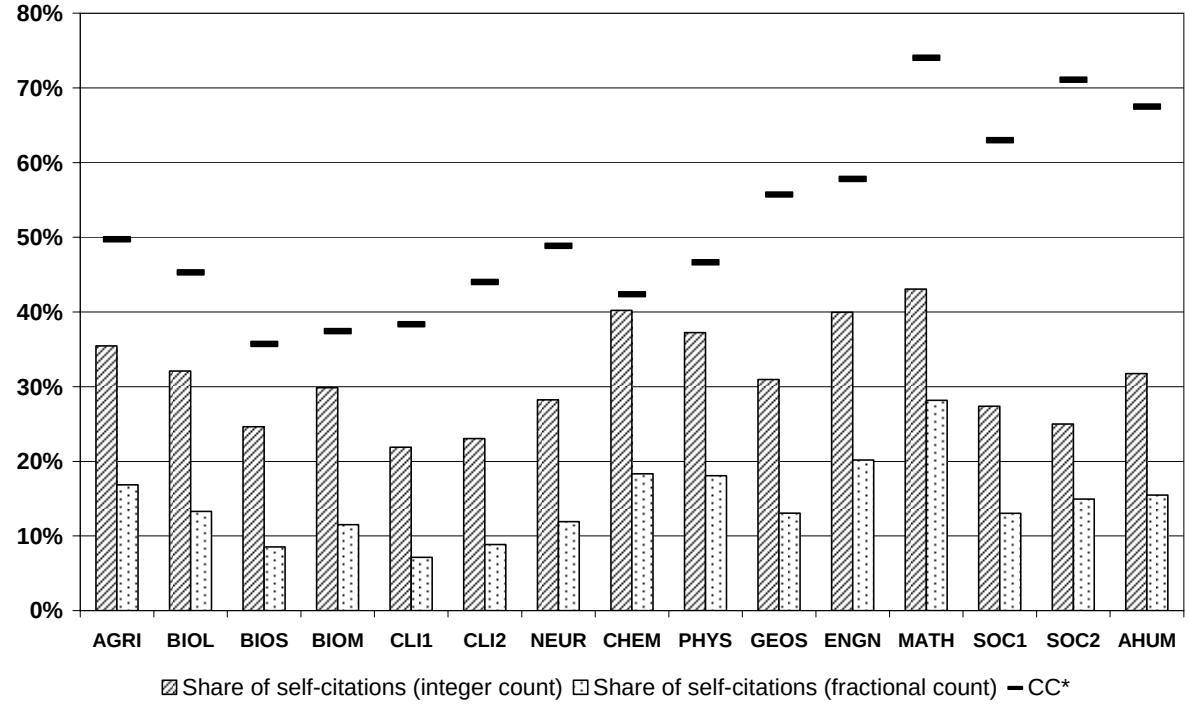

Figure 3 Share of binary and fractional self-citations in all citations in 15 different subject areas of the sciences, social sciences and humanities (subject codes can be found in the Appendix)

National statistics on author self-citations in all science fields combined are given in Table 1. The 40 most active countries in 1994 have been selected; this corresponds to a threshold of 1500 publication indexed in the Science Citation Indexed Expanded (SCIE). Here we excluded social sciences and humanities. Citation statistics have been calculated for a 3-year and 10-year citation window, respectively. Countries have been ranked by fractional self-citation share (according to Jaccard Index) in 1994-1996 in descending order. The results for the two periods can hardly be explained with language-specific counting errors.

Beyond any doubt, national publication profiles may certainly have strong influence through subject-specific peculiarities (cf. Figure 2). However, also different profiles cannot explain the large range of fractional shares (>25% in Ukraine and Slovakia and < 10% in Switzerland and USA in 1994-1996) among the countries alone. Thus we find highly developed countries at the bottom of the list regardless of the language spoken in these countries. Also the different ageing of self-citations in several countries is striking (cf. Singapore and Bulgaria, Brazil and Hong Kong in Table 1).

In order to gain a deeper insight in the national characteristics, subject profiles and research performance in the context of self-citations, a sample of 60 institutions representing six European countries have been selected. The sixty most active universities, research institutes and companies from Austria, Denmark, Finland, Hungary, Ireland and Sweden have been selected. This selection corresponds to a threshold of 144 publications indexed in the 2000 volume of the WoS. The most active institutions in the selection are publishing far above 2000 papers per year. The six countries are, of course, not evenly presented. Finland is represented by the largest number of institutions, followed by Sweden and Hungary. Ireland, the "smallest" country in the set, is represented by just two institutions.

Table 1 Share of binary and fractional self-citations in all citations of the 40 most active countries in the sciences (ranked by fractional self-citation share in 1994-1996 in descending order)

Rank	Country	Share of self-citations (1994-1996)		Share of self-citations (1994-2003)	
		Binary	Fractional	Binary	Fractional
1	Ukraine	55.7%	27.6%	42.1%	17.1%
2	Slovakia	52.6%	25.4%	39.0%	15.1%
3	Egypt	49.7%	24.6%	32.4%	13.4%
4	India	45.1%	23.3%	32.1%	13.2%
5	Russia	47.5%	22.2%	36.6%	14.5%
6	Poland	47.4%	22.0%	35.3%	13.7%
7	Czech Republic	47.9%	21.0%	34.6%	12.7%
8	P.R. China	46.0%	20.9%	32.5%	11.5%
9	Singapore	40.7%	20.6%	23.8%	9.4%
10	Bulgaria	44.6%	20.5%	31.0%	12.1%
11	Taiwan	41.2%	18.9%	25.9%	9.6%
12	Turkey	41.7%	18.6%	28.3%	10.1%
13	Greece	43.6%	18.5%	29.0%	10.2%
14	Korea	40.9%	17.3%	28.6%	9.7%
15	Spain	39.8%	17.1%	28.3%	9.9%
16	Hungary	39.7%	16.7%	29.9%	10.2%
17	Argentina	38.9%	16.5%	27.7%	9.9%
18	South Africa	34.4%	15.5%	22.7%	8.6%
19	Japan	35.5%	15.4%	24.4%	8.2%
20	Mexico	37.8%	15.4%	26.9%	9.0%
21	Brazil	39.4%	15.2%	30.2%	9.5%
22	Hong Kong	35.0%	14.8%	22.0%	7.1%
23	New Zealand	31.7%	13.8%	19.5%	6.9%
24	Australia	30.9%	13.2%	19.9%	6.9%
25	Italy	34.3%	13.1%	23.6%	7.3%
26	Norway	32.1%	12.9%	19.9%	6.4%
27	Austria	34.6%	12.8%	22.9%	6.8%
28	Israel	30.7%	12.6%	21.3%	7.2%
29	Germany	33.2%	12.4%	22.5%	6.8%
30	Denmark	33.4%	12.2%	21.4%	6.3%
31	Sweden	32.2%	12.0%	21.0%	6.3%
32	France	32.1%	11.6%	22.3%	6.5%
33	Belgium	33.0%	11.6%	22.1%	6.2%
34	Ireland	30.3%	11.6%	19.0%	5.9%
35	Canada	29.0%	11.4%	19.0%	6.1%
36	Finland	31.3%	11.4%	20.7%	6.1%
37	Netherlands	30.8%	10.9%	20.1%	5.7%
38	UK	28.3%	10.8%	18.4%	5.7%
39	Switzerland	27.0%	9.2%	18.5%	5.1%
40	USA	23.8%	8.8%	15.7%	4.7%

Most of the selected European institutions are universities; among those there are some of them are specialised one, such as medical and technical universities. Besides the institutions of higher educations, non-university academic institutes and companies are represented.

Hierarchical clustering with squared Euclidean distances and Ward-linkage was used to create clusters of likewise institutes in terms of publication profile on basis of the publication output in 2000 and 2001. Selected universities could be thus assigned to the following three profiles in 2000: Class 1 with predominant natural and engineering sciences, Class 2 with focus on Biological and applied biological fields; agriculture and Class 3 with main focus on medical research. The results are presented in Table 2; the institutes are treated anonymously.

Table 2 Share of binary and fractional self-citations in all citations of the 60 most active institutes in six selected European countries in 2000 (ranked by fractional self-citation share in 2000 in descending order)

#	Country	Class	Papers	Share of self-citations		#	Country	Class	Papers	Share of self-citations	
				Binary	Fractional					Binary	Fractional
1	H	1	164	50.0%	23.3%	31	A	1	249	32.9%	11.2%
2	H	1	252	50.3%	22.0%	32	H	3	441	31.3%	10.8%
3	H	2	169	47.5%	19.1%	33	S	1	780	28.1%	10.6%
4	FIN	1	193	46.5%	18.6%	34	FIN	1	720	32.2%	10.5%
5	FIN	1	625	43.1%	17.7%	35	S	3	920	31.7%	10.4%
6	S	1	1043	40.8%	17.0%	36	H	3	229	28.7%	10.3%
7	H	1	387	39.6%	16.6%	37	S	3	2229	29.7%	10.2%
8	S	1	193	37.1%	16.3%	38	A	3	939	30.0%	10.2%
9	H	1	443	39.8%	16.0%	39	S	3	1821	28.9%	10.1%
10	FIN	1	245	40.2%	15.9%	40	A	3	823	28.3%	9.8%
11	FIN	1	192	40.0%	15.5%	41	FIN	3	1018	29.3%	9.7%
12	H	1	460	41.4%	15.0%	42	DK	3	1493	28.7%	9.4%
13	A	1	311	38.2%	14.9%	43	DK	3	2476	28.3%	9.0%
14	DK	1	337	40.5%	14.8%	44	S	3	2601	27.8%	8.9%
15	FIN	3	417	39.8%	14.7%	45	A	3	2410	28.0%	8.8%
16	DK	1	233	40.4%	14.7%	46	S	3	938	26.3%	8.7%
17	A	1	274	36.5%	14.5%	47	IRE	3	626	26.1%	8.5%
18	S	1	908	35.9%	14.5%	48	FIN	3	2337	27.3%	8.2%
19	H	1	490	38.9%	14.3%	49	FIN	3	460	26.6%	8.1%
20	DK	2	240	38.4%	14.2%	50	FIN	3	354	23.9%	7.8%
21	A	1	745	35.5%	13.8%	51	FIN	3	388	24.4%	7.7%
22	H	3	524	40.4%	13.7%	52	DK	3	646	25.3%	7.6%
23	DK	1	767	35.2%	13.3%	53	FIN	3	151	23.7%	7.5%
24	FIN	1	232	30.5%	12.4%	54	FIN	3	333	22.4%	7.3%
25	A	2	211	34.4%	12.2%	55	DK	3	293	24.1%	6.8%
26	DK	2	516	35.5%	12.2%	56	S	3	2703	22.8%	6.7%
27	H	3	144	26.9%	12.1%	57	DK	3	204	26.1%	6.7%
28	A	2	182	34.0%	12.1%	58	S	3	274	20.1%	6.5%
29	S	2	785	34.1%	12.0%	59	FIN	3	442	21.9%	5.7%
30	IRE	3	1434	29.8%	11.9%	60	FIN	3	175	19.3%	4.8%

No doubt, national characteristics can be found in Table 2 as well as the influence of subject profiles. However, the shares of self-citations of the institutions 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.

Also the range of self-citation shares is much larger than in the national case (cf. Table 1). This observation is in line with results of an earlier study by the authors (see *Thijs* and *Glänzel*, 2005). Class 1 institutions usually have high shares of fractional self-citations while Class 3 institutions can be found rather at the bottom. Although institution ranking according to the two self-citations shares seems by and large to coincide, the indicator values of several institutes (for instance, #8 and #23) are striking exceptions to this rule. The decrease of self-citation shares through fractionating is just as impressive as in the macro case.

Conclusions

The most striking feature of fractional self-citation counting is the extremely fast ageing. Three years after publication the Jaccard-based measure indicates a self-citation share of 15% for the world total; after ten years this share amounts to 9%. Author self-citations become after such a long period practical negligible. However, also three years after publication, the effect of author self-citations is quite low if compared with binary counts.

The reason why self-citations so rapidly lose their weight as time elapses might be interpreted in the light of the following assumptions. Co-authorship and, above all, multi-authorship is not merely the results of the work of stable teams, that is, multi-authorship might also be caused by occasional collaboration of one or more of co-authors who are not continuing research in the mainstream of the team they collaborated with. Also the constitution of the stable kernel of the research teams might considerably change over periods of five-ten years. Moreover, the increasing mobility of scientists might have a strong effect in this context, too. Finally, scientists and their collaborators are obviously more interested in continuing and applying most recent work while other colleagues outside their teams use scientific information published in their papers still a long time after the research projects have been completed.

References

Ajiferuke, I, Burrell, Q & Tague, J (1988), Collaborative coefficient: A single measure of the degree of collaboration in research. *Scientometrics*, 14, 421-433.

Aksnes, D W, (2003), A macro-study of self-citations. *Scientometrics*, 56, 235-246.

Glänzel, W & Schubert, A, (2003), A new classification scheme of science fields and subfields designed for bibliometric evaluation purposes, *Scientometrics*, 56, 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), Does co-authorship inflate the share of self-citations? *Scientometrics*, 61, 395-404.

Nederhof AJ, Meijer RF, Moed HF, van Raan AFJ (1993), Research Performance Indicators for University Departments - A Study of an Agricultural University, *Scientometrics*, 27, 157–178.

Snyder, H & Bonzi, S (1998), Patterns of self-citation across disciplines. *Journal of Information Science*, 24, 431-435.

McNee, S, Albert, I, Cosley, D, Gopalkrishnan, P, Lam, SK, Rashid, AM, Konstan, JA & Riedl, J (2002), On the Recommending of Citations for Research Papers. In Proceedings of ACM 2002 Conference on Computer Supported Cooperative Work (CSCW2002), New Orleans, LA, pp. 116-125.

Thijs, B & Glänzel, W (2004), ***The influence of author self-citations on bibliometric meso-indicators. The case of European universities.*** **Paper presented at the 8th Conference on S&T Indicators, Leiden (Netherlands), 24 September 2004.**

Appendix

Key to field abbreviations and subject codes

Abbreviation	Code	Subject Field
AGRI	A	Agriculture & Environment
BIOL	Z	Biology (Organismic & Supraorganismic Level)
BIOS	B	Biosciences (General, Cellular & Subcellular Biology; Genetics)
BIOM	R	Biomedical Research
CLI1	I	Clinical and Experimental Medicine I (General & Internal Medicine)
CLI2	M	Clinical and Experimental Medicine Ii (Non-Internal Medicine Specialties)
NEUR	N	Neuroscience & Behaviour
CHEM	C	Chemistry
PHYS	P	Physics
GEOS	G	Geosciences & Space Sciences
ENGN	E	Engineering
MATH	H	Mathematics
SOC1	S	Social Sciences I (General, Regional & Community Issues)
SOC2	O	Social Sciences II (Economical & Political Issues)
AHUM	U	Arts & Humanities