There goes another one: Introducing the NUCA-set of indicators

Miloš Jovanović[1,2], Frank Fritsche[1]

[1]*milos.jovanovic@int.fraunhofer.de*
[1]Fraunhofer INT, Appelsgarten 2, 53879 Euskirchen (Germany)
[2]Heinrich-Heine-University, Universitätsstr. 1, 40225 Düsseldorf (Germany)

frank.fritsche@int.fraunhofer.de
[1]Fraunhofer INT, Appelsgarten 2, 53879 Euskirchen (Germany)

Abstract
Indicators that provide a means of evaluation of research have been at the heart of bibliometric studies since the beginning of the discipline. In this research-in-progress paper we introduce a new set of measures and indicators in order to complement existing evaluation methods of research excellence for single authors and highlight a new facet of bibliometric methods. Existing measures normally use publication and citation numbers as the basis for the calculation of an indicator. The NUCA-set (Number of unique citing authors) of indicators uses the number of authors that cited the evaluated author. These new indicators visualize cases where an author's work might have accumulated many citations but in reality only has been frequently cited by a small number of people. The NUCA-set leads to a more balanced evaluation and a more comprehensive evaluation profile if used in addition to already existing indicators like the H-index or the citation rate.

Introduction
A large part of the discipline of bibliometrics deals with the design and calculation of indicators. These are used to evaluate various entities, for example a country, a journal, an institution or a single author. In this research-in-progress paper we propose a new set of indicators that allows a more comprehensive evaluation of single authors. To do this, we do not focus on the number of citations an author has received. Rather, we analyse the number of authors that cited the evaluated author's publications.
Over the years, several bibliometric indicators have been created for various reasons, e.g. to figure out which journals should be subscribed by libraries or how old the cited literature of different scientific disciplines is. For a more detailed look at the different bibliometric indicators we refer the reader to some extensive reviews on the subject (e.g. Wallin, 2005; Bar-Ilan, 2008; Thompson, Callen & Nahata, 2008; Alonso et al., 2009). Some of these indicators have become very popular and are being used in various contexts. Among the most famous bibliometric indicators are the impact factor for journals, the citation rate, and H-Index for authors (Garfield & Sher, 1963; Vinkler, 2005; Hirsch, 2005).
Especially the indicators that are being used to evaluate scientists are regularly discussed and criticised. This is due to the fact that these indicators allow the creation of rankings in which an author at the top is perceived as being "better" than the ones further down the ranking. For example, author self-citation, its employment in bibliometric analyses and its influence on such a ranking of authors are topics which are discussed in various papers (Aksnes, 2003; Fowler & Aksnes, 2007). Another issue in this context is the influence of biased citing, which refers to the possibility that an author cites other authors not because he used their work but because of other reasons, like, for example, the cited author being a colleague of his (MacRoberts & MacRoberts, 1996; Bornmann & Daniel, 2008). The question on how to deal with citations to papers with multiple authors is also raised, since it is possible to use total or fractional counting. Depending on whether one uses the one variation of counting or the other can change the sequence of a ranking (Gauffriau & Larsen, 2005). One answer to this criticism of citation indicators is to use a set of indicators and not just one of them in order to

provide a more comprehensive impression of an evaluated author (Franceschini, Galetto & Maisano, 2007). With this paper, we introduce new indicators that can complement such a set of indicators.
Most bibliometric indicators that are used to evaluate a single author involve the number of citations such an author has received. This is often combined with the number of papers he has published. But behind every citation lies at least one author who has created this citation by including a paper into the reference list of his work. We are of the opinion that it makes sense to take into account the number of authors that lie behind each citation to an evaluated author's work. By doing so we are convinced that some of the above criticism on indicators and citation analysis in general can be alleviated. There exists at least one other indicator that also takes the number of citing authors into account, the "Science Impact Index". This indicator however is being counted differently and does not take the same data into account as our new set of indicators (Lehrl, Kinzel & Fisch, 1988; Lehrl, 2005).

Method

In order to demonstrate the effectiveness of this new aspect we counted a number of different measures with which we calculated our new set of indicators. The entities under evaluation are individual authors. Thus, the first data set that is needed for the evaluation is a complete set of the author's bibliography. Ideally, this is supplied by the author himself but it can also be compiled by searching in a citation data base. We use employed the Web of Science, accessed via the Fraunhofer-Gesellschaft. This access encompasses the Science Citation Index Expanded, the Social Science Citation Index (both for the years 1970-present), the Conference Proceedings Citation Index – Science, and the Conference Proceedings Citation Index – Social Science & Humanities (both for the years 1990-present). The evaluated author's bibliography is then downloaded for further analysis. In this data set (we call it the "author data set") we count N_p, which is the number of papers the evaluated author has published.
In a second step, we determine the papers citing the author data set. This is due to the fact that the author data set only includes the number of times a publication has been cited but not exactly by which papers. For this, we used the "Citation report"-tool of the Web of Science. With it, we were able to download a second data set (the "citation data set") comprising all publications that cite the work of the author under consideration.
Next we had to determine exactly which publications from the citation data set cited which papers from the author data set. For this allocation we wrote a Perl-script to pair the citing publications to the cited publications. The script analyses the "Cited references"-field in the citation data set to find the respective paper in the author data set. Each reference could be split into first author, journal, volume, first page, article number, public year and Digital Object Identifier (DOI). We take any information and match it with the proper field in the author data set. Unfortunately, not every reference has all the information. With a DOI, we have an easy match. Without it, though, we must find a paper in the author data set which matches all the fields of the reference. In addition, there must be a first author and a journal name. We use a similarity algorithm and a threshold to identify the journal name. For the authors that we evaluated for this contribution our method worked well. We were even able to pair papers that were not paired in the Web of Science. Why some of the papers were not paired correctly in the Web of Science was not always clear in the cases where the data was indexed correctly in the data base. For now, it seems as if these citations were simply overlooked.
But we are aware of some shortcomings still present in our script. For example, few references in the citation data set contain combined data like the volume and issue number together. In the author data set this data is included separately and thus could not yet be

identified by our script. However, this was rarely the case. For most publications, the algorithm we employed worked well. In some cases we had to manually correct data in order to accurately pair publications. In this respect, papers from the new conference proceedings citation indices were most critical, as the names of the sources show many variations in their abbreviation. Because of this we introduced an algorithm into our script which used a threshold value for the allowed similarity of source names in the author and citation data set. Depending on the data set, this value has to be adjusted to obtain the best result possible. For example, a low threshold value will identify an article that has been published in the journal "Signal Processing" and pair it with a proceedings paper from the "Proceedings of the European Signal Processing Conference". In order to avoid this, the threshold needs to be raised high enough to not identify such incorrect pairs but to be tolerant enough to pair papers in which the source has simply been misspelled. Our results show that the number of citations we found and allocated is about 5% higher than the one allocated by the Web of Science.

Once the pairing of the data sets was completed we calculated the number of citing authors *NCA*. This is the number of all authors of the citation data set. This includes authors that repeatedly cited the same paper from the author data set. Thus, if a "Smith, J" cited a paper from the author data set ten times then his name would also count ten times in the *NCA*.

Next we counted the number of unique citing authors *NUCA*. For this measure we counted every author citing a certain publication in the author data set only once. The reason for this is that in our opinion it makes sense that a researcher cites another author's work because he used that work in his own research. However, if this researcher cites the same publication over and over again it seems unlikely that he learns anything new from it. However, it is still possible for this researcher to cite different works in the author data set. The *NUCA* thus represents the measure of how many people used the works of an author. We are aware of the problem with homonyms. If three different people which are all named "Smith, J" cited a publication then they would be treated as one person. This problem might be dealt with by taking an author's affiliation into account. However, author's that change their working place often would still pose a problem. The *NUCA* measure is similar to the above mentioned "Science Impact Index".

The *NCA* and *NUCA* values were then divided by the number of publications in the author data set. This leads to the NCA- and NUCA-Factor, which are essentially an average number of citing authors per paper:

$$NCA-Factor=\frac{NCA}{N_p} \tag{1}$$

$$NUCA-Factor=\frac{NUCA}{N_p} \tag{2}$$

In a second step we removed those publications from the denominator N_p that were not cited at all. We called this measure N_U. This leads to the NCA-Factor$_U$ and NUCA-Factor$_U$ indicators:

$$NCA-Factor_U=\frac{NCA}{N_U} \tag{3}$$

$$NUCA-Factor_U=\frac{NUCA}{N_U} \tag{4}$$

Finally, we calculated the NUCA-Index. Inspired by the H-Index, an author has a NUCA-Index of *n* if he has *n* papers which have been cited by *n* unique authors.
With this set of five new indicators, it is possible to give a more comprehensive citation profile of an author.
We would like to illustrate this with two stereotypical examples before we present actual results: First, let us consider an author who has published 10 papers and only cites himself in each of these papers. In this case he would have 45 citations, a citation rate of 4,5 and an H-Index of 5. These values are not unusual and thus it would not be clear that the author has only cited himself over the years. However, the value of the NUCA-Index would be 1. For the other four indicators the value would be 0,1 (see equations 1 to 4). These numbers show that the citations have been made by only one person. Of course, ignoring self-citations would also lead to a similar result. But self-citation can be a legitimate mean if one has to cite one's own work because nobody else has worked in a given specific field. Thus, we think it is more adequate to consider self-citations in a fair manner. Ignoring self-citations also will not help in the second example.
For this example let us suppose that there is an author with 10 publications, each cited ten times. However, these 100 citations have been made by a group of ten colleagues of his, each citing every publication once. The author has never cited himself. This author thus has a citation rate of 10 and an H-Index of 10. For this author, the NUCA-Index would also have a value of 10, showing that all of his papers were cited by ten authors. But the other indicators would have a value of 1. This reveals that the citations have been made by a small number of people.

Results

For a first real-life test of the set of new indicators, we conducted a small case study by choosing four eminent researchers from the Fraunhofer-Gesellschaft. All of them are from different institutes and different scientific disciplines. We did this in order to check, whether our new indicators would produce irregular results.
In table 1, we show the bibliometric measures for these four authors:

Table 1. Bibliometric measures of four researchers from the Fraunhofer-Gesellschaft.

Author	*Pub.*	*Cit.*	*NCA*	*NUCA*
A	308	6904	33875	20938
B	376	7575	44340	27663
C	338	923	2435	1446
D	66	118	392	307

For each author we calculated the citation rate, H-Index, and all of the new indicators. Results are summarized in Table 2:

Table 2. Results of the bibliometric evaluation of four researchers from the Fraunhofer-Gesellschaft using standard indicators and the new set of NUCA-Indicators.

Author	*Citation rate*	*H-Index*	*NCA-Factor*	*NCA-Factor$_U$*	*NUCA-Factor*	*NUCA-Factor$_U$*	*NUCA-Index*
A	22,42	43	109,98	146,01	67,98	90,25	75
B	20,15	40	117,93	148,29	73,57	92,52	86
C	2,73	13	7,2	13,9	4,27	13,9	16
D	1,79	6	5,94	13,52	4,65	10,59	10

Compiling a sorted list from the various indicators leads to the following rankings:

1. Citation rate: A B C D
2. H-Index: A B C D

3. NCA-Factor: B A C D
4. NCA-Factor$_U$: B A C D
5. NUCA-Factor: B A D C
6. NUCA-Factor$_U$: B A C D
7. NUCA-Index: B A C D

One can already see with these lists that a ranking based on the NUCA-indicators will lead to slightly different results than H-Index and citation rate rankings. The fact that authors C and D switch places in the ranking based on the NUCA-Factor but not in the one based on the NUCA-Factor$_U$ suggests that author D has more citing persons per paper (NCA-Factor) than author C but also more papers that have been cited at least once. However, there is one more interesting aspect about the NUCA-set of indicators. While the difference in the H-Indices of authors C and D are quite high (more than 100%) this discrepancy cannot be found in the NUCA-indicators. In general, the difference of the new indicators in the rankings is not as pronounced as with the citation rate and the H-Index.
Apart from the numbers of citations and publications, the NUCA-Indicators give a better idea of how many people possibly have perceived the work of an author. The similarity in ranking suggests that the citation rate and H-Index might be correlated to the new indicators. However, cases of extensive self-citation or citation bias would produce different values as shown with the two stereotypical examples above.

Discussion

In this paper, we introduced a new set of indicators that in our opinion leads to a more comprehensive evaluation of an author's published work. The main advantages of using these new indicators lie in their possibility to disclose extensive self-citation and biased citation.
Based on our first results, we recommend considering the complete new set of indicators for the purpose of evaluating an author. One of the reasons for this recommendation is implied by the calculation of the indicators itself. If, for example, one only considers the NUCA-Factor$_U$ for the purpose of evaluating a group of authors, scenarios are possible where a higher NUCA-Factor$_U$ does not give a sound estimate of how many people on average have perceived an author's publications. Consider one author with only one paper that has been cited very often by different persons. Compared to another author with many papers that in total have been cited just as often by the same people, this first author would have a much higher NUCA-Factor$_U$, even though both have similar measures and the same NUCA-Factor. Another reason to use the complete set is a problem similar to the question of total and fractional counting. Here, one could criticise that a paper with 100 authors counts much more than a paper with a single author if calculated with the NUCA-indicators.

Conclusion

Our results show that the new set of indicators gives a more differentiated impression of an author's published work. In future studies, we will elaborate on the question of the performance of the new indicators in the different scientific disciplines. It is quite obvious, that a comparison of authors from different scientific disciplines is problematic, since in some disciplines like clinical medicine or physics there exist more cases of collaboration than in

mathematics or agriculture (Melin & Persson, 1996). With a larger group of evaluated authors we will also analyse the correlation between standard bibliometric indicators and the new set of indicators and the course of an author's NUCA-Indicators over time. It will also be interesting to apply the NUCA-Indicators to authors which have been known to cite themselves extensively or are being cited very often by the same people without any comprehensible reason in order to see what results this will produce in comparison with standard indicators.

References

Aksnes, D. W. (2003). A macro study of self-citation. *Scientometrics*, 56 (2), 235-246.

Alonso, S., Cabrerizo, F. J., Herrera-Viedma, E., Herrera, F. (2009). *Journal of Informetrics*, 3 (4), 273-289.

Bornmann, L. & Daniel H.-D. (2008). What do citation counts measure? A review of studies on citing Behavior. *Journal of Documentation*, 64 (1), 45-80.

Bar-Ilan, J. (2008). Informetrics at the beginning of the 21st century—A review. *Journal of Informetrics*, 2, 1-52.

Fowler, J. H. & Aksnes, D. W. (2007). Does self-citation pay? *Scientometrics*, 72 (3), 427-437.

Franceschini, F., Galetto, M., & Maisano, D. (2007). *Management by measurement: Designing key indicators and performance measurement systems*. Berlin: Springer Verlag.

Garfield, E., Sher, I. H. (1963). New factors in the evaluation of scientific literature through citation indexing. *American Documentation*, 14 (3), 195–201.

Gauffriau, M. & Larsen, P. O. (2005). Counting methods are decisive for rankings based on publication and citation studies. *Scientometrics*, 64 (1), 85-93.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences*, 102, 16569-16572.

Lehrl, S., Kinzel, W., & Fischer, B. (1988). Der Science Impact Index. Untersucht an den Ordinarien der bundesdeutschen Psychiatrie und Neurologie. In: Daniel, H.-D. & Fisch, R. *Evaluation von Forschung: Methoden, Ergebnisse, Stellungnahmen.* (pp. 291-305). Konstanz: Universitätsverlag.

Lehrl, S. (2005). Längsschnittuntersuchung zur Zitierhäufigkeit Gelehrter in der HNO-Heilkunde. *HNO*, 5, 415-422.

MacRoberts, M. H. & MacRoberts, B. R. (1996). Problems of citation analysis. *Scientometrics*, 36 (3), 435-444.

Melin, G. & Persson, O. (1996). Studying research collaboration using co-authorships. *Scientometrics*, 36 (3), 363-377.

Thompson, D. F., Callen, E. C., Nahata, M. C. (2008). New Indices in Scholarship Assessment. *American Journal of Pharmaceutical Education*, 73 (6), 1-5

Vinkler, P. (2005). Eminence of scientists in the light of the h-index and other scientometric indicators. *Journal of Information Science*, 33 (4), 481-491

Wallin, J. A. (2005). BibliometricMethods: Pitfalls and Possibilities. *Pharmacology & Toxicology*, 97, 261-275.