# CITATION CLASSES[1]:
# A NOVEL INDICATOR BASE TO CLASSIFY SCIENTIFIC OUTPUT

Wolfgang Glänzel[*], Koenraad Debackere[**], Bart Thijs[****]

[*] *Wolfgang.Glänzel@kuleuven.be*
Centre for R&D Monitoring (ECOOM) and Dept. MSI, KU Leuven, Leuven (Belgium)
Department of Science Policy & Scientometrics, Library of the Hungarian Academy of Sciences, Budapest (Hungary)

[**] *Koenraad.Debackere@kuleuven.be*
Centre for R&D Monitoring (ECOOM) and Dept. MSI, KU Leuven, Leuven (Belgium)

[***] *Bart.Thijs@kuleuven.be*
Centre for R&D Monitoring (ECOOM) and Dept. MSI, KU Leuven, Leuven (Belgium)

## Abstract

In the present paper we present use cases and scenarios for a new assessment tool based on bibliometric data and indicators. The method, the application of which is illustrated at different levels of aggregation (national, institutional and individual researchers), aims at building on traditional indicators while stepping away from the linear ranking model and thus at providing enough space for a more differentiated analysis of published research results. The method proved statistically robust and insensitive to the influence of the otherwise commonly encountered database-related, communication- and subject-specific factors. This finding is a clear step forward in the use of bibliometric data for assessment and evaluation purposes, as often is a need of science policy.

## 1. Introduction: The need for multi-level profiling of excellence

### *The conceptual background*

One of the objectives of previous studies (Glänzel, 2013a,b) was to analyse to what extent the high-end of citation impact as reflected by the tail of scientometric distributions is in line with the "standard" citation impact attracted by the majority of scientific papers and to what extent 'outliers' might be responsible for possible deviations and distortions of citation indicators. Two important observations are relevant in this context. One solution proposed in these studies was to use tail indices as a supplement to traditional citation-based performance indicators, such as the share of uncited papers to the mean citation rate. The analysis of the tail, which was based on ordered or ranked observations, can practically be uncoupled from the overwhelming rest of the empirical distribution. Most studies of the tail of scientometric distributions proceed from a Pareto model. The estimation of the tail parameter can directly be obtained from subsets of order statistics. This approach even allows one to construct confidence intervals for its estimator. Nevertheless, the estimation of the tail index remains rather problematic since most methods are still sensitive to the cut-off point for the tail. Since already minute changes of the tail parameter might have significant consequences in an evaluative context, the recommendation in the study by Glänzel (2013a) was to favour a parameter-free solution for the assessment of outstanding performance. This might also help avoid parameter conflicts resulting from estimating parameters on the basis of head and trunk

[1] All data presented in the tables and figures of this study are based on data sourced from Thomson Reuters Web of Science Core Collection.

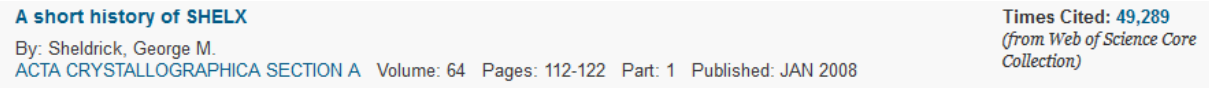
of the distributions, on one hand, and from their tail, on the other hand. Therefore, a "reduction" of the original citation distribution to performance classes on the basis of *Characteristic Scores and Scales* (CSS) introduced by Glänzel & Schubert (1988) was proposed as an alternative parameter-free solution.

## 2. Application: Profiling excellence in practice

*General features and properties*

The main advantage of this method and its superiority over other methods aiming at providing performance classes certainly lies in the two above-mentioned basic properties: CSS is parameter-free, that is, the method is not error-prone regarding estimation methods and random errors of resulting parameters, and, secondly, it is self-adjusting, that is, no pre-defined thresholds, such as given numbers of citation or percentiles, are necessary. Unlike in methods based on percentiles (e.g., Leydesdorff et al. (2011), this approach is not sensitive to ties and ensures seamless integration of measures of outstanding and even extreme performance into the standard tools of scientometric performance assessment. The method can be interpreted as a reduction of the original citation distribution to a distribution over a given number of performance classes. The number of profile classes used for the evaluation is the only "arbitrary" number that needs to be defined in the beginning. Usually four classes are sufficient. The four classes stand for 'poorly cited' (Class 1), 'fairly cited' (Class 2), 'remarkably cited' (Class 3) and 'outstandingly cited' (Class 4) papers. Papers in class 3 and 4 can be considered highly cited.

A further advantage, that this method shares with other performance-class approaches, e.g. based on percentiles, is its insensitivity for outliers: An extreme citation rate becomes just one out of several highly cited publication assigned to the highest performance class. This prevents outliers from biasing or even distorting indicators of the unit under study. It is clear that extreme citation rates as the ones shown in Figure 1 would distort most traditional citation indicators (based on mean values) even for larger institutions. The paper on the history of SHELX was published by a researcher of the University Göttingen. The extreme citation impact of this single paper is able to distort world university rankings if this is based on citation counts or means. This effect was first observed by Waltman et al. (2012). We present the up-to-date citation impact of this paper just as an illustration (Figure 1).



A short history of SHELX

By: Sheldrick, George M.

ACTA CRYSTALLOGRAPHICA SECTION A Volume: 64 Pages: 112-122 Part: 1 Published: JAN 2008

Times Cited: 49,289
*(from Web of Science Core Collection)*

*Figure 1 Extreme citation rate received by one single paper published in 2008 till 15 July 2016 [Data sourced from Thomson Reuters Web of Science Core Collection]*

The CSS method proceeds from a baseline model that usually consists of the complete publication "universe", that is, a complete bibliographic database, or more precisely, a complete citation index. In our study we used Thomson Reuters *Web of Science Core Collection* but we stress that Elsevier's Scopus could be used as well.

In a nutshell, characteristic scores are obtained by iteratively calculating the mean value of a citation distribution and subsequently truncating this distribution by removing all papers with less citations than the conditional mean. In order to obtain fours classes, the process is stopped after three iterations. Although the method is based on mean values, the huge number of publications underlying the base-line guarantees that the influence of outliers remains marginal. A further issue in comparative assessment at practically any level of aggregation arises from the peculiarities of scholars' communication behaviour in their research domains and disciplines. This results in different disciplinary standards of publication activity and citation impact. In order to be able to apply the method to the assessment of national research output and of multidisciplinary universities, one has to apply kind of "fractional" normalisation, when calculating the threshold values for class assignment. The detailed description of the procedure and its mathematical background is documented in previous studies by the authors (e.g., Glänzel, 2013b; Glänzel et al., 2014).

*The national and institutional level*

Taking into account that citation standards considerably differ in the various disciplines, across countries, regions and institution, the method was developed for benchmarking and the comparative assessment of research performance at the *national level, for institutional entities and individual scientists*. The performance classes obtained from this method can be applied to the comparative analysis of the citation-impact profiles of given units amongst themselves as well as with the reference standard in the given subject. It has been stressed that the calculation of a "single" indicator over these classes is not suitable as this would reduce the gained added value and thus destroy the advantages of the method. However, it has also been shown that the application to combinations of different disciplines is indeed be possible Glänzel (2013b). The study has furthermore demonstrated robustness of the method for combinations of disciplines with respect to the publication year and the citation window at the level of national and institutional research performance. In the first empirical paper (Glänzel et al., 2014) we first extended the analysis to a specific level of aggregation, particularly to the assessment of research institutions. At this level the number of publications per unit is considerably lower than at the national level but more important is that we expected to observe more diverse research profiles. In particular, some institutions have a specialised profile while others are truly multidisciplinary in their research activities. The aim of the first part of the paper is therefore to demonstrate the robustness of the method at the meso level and its independence of institutional publication profiles.

Four major issues proved challenges to bibliometric tools in terms of stability and robustness. The first one has already been mentioned: the subject-specific scholarly communication behaviour and the resulting disciplinary standards. The next two issues are more straightforward: scientists are able to produce more results and to write more papers in a longer period that in a shorter one and, similarly, papers can accumulate more citations during, say, 10 years than during only two or three years. The last phenomenon is, though less obvious, also of high relevance: communication behaviour of scientists in the same field might change over time and also the data sources used for evaluation are subject to evolutionary and structural changes. This phenomenon and its effect on bibliometric indicators has already been studied, among others, by Persson et al. (2004).

The following example uses two different publication years along with two different citations windows to illustrate the method's robustness. All data have been extracted and processed from bibliographic rough data of the Web of Science Core Collection. Figure 1 displays national shares in the four citation classes for six selected countries. Although CSS is not directly linked to percentiles, the baseline provides a distribution of papers over classes of about 70% (Class 1), 21% (Class 2), 6%–7% (Class 3) and 2%–3% (Class 4), independently of publication year and citation window. Also Albarrán and Ruiz-Castillo (2010) found a similar 70–21–9 rule when they combined the two upper performance classes for papers published between 1998 and 2002 and using a 5-year citation window. Further evidence of robustness, subject and time invariance can be found in Glänzel (2007).

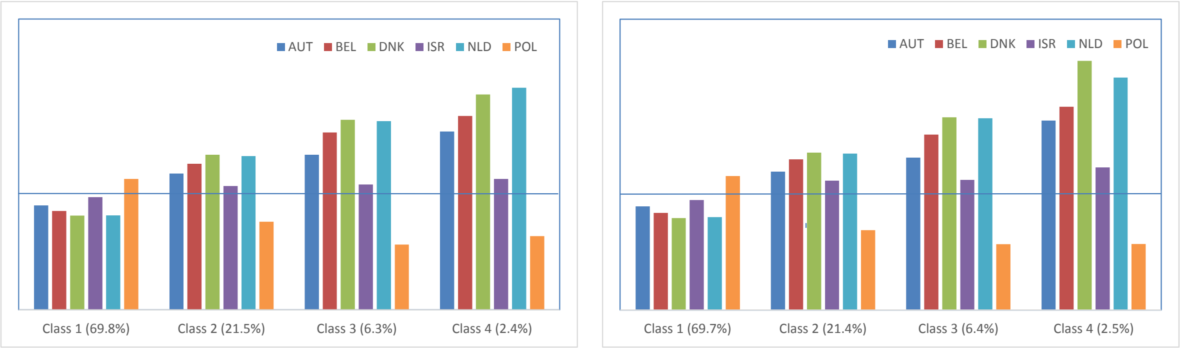


*Figure 2. National shares of publications in the four CSS classes in all fields combined in 2007 with 5-year citation window (left) and 2009 with 3-year citation window (right); the world standard is indicated by the horizontal line. [Data sourced from Thomson Reuters Web of Science Core Collection]*

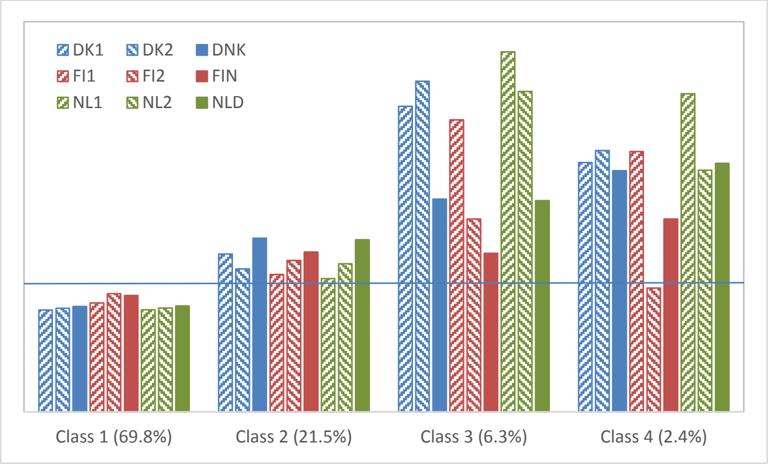


*Figure 3. Shares of publications of selected universities and countries in all fields combined in 2007 (5-year citation window) [Data sourced from Thomson Reuters Web of Knowledge]*

Figure 3, finally, shows the institutional comparison of class profiles for which we have selected two European universities each. Although the universities' profiles mostly mirror the national patterns, we find also situations different from the respective national standard. Thus we find a less favourable situation for the university IT1 in Italy and FI2 in Finland. The high standard of the selected Danish and Dutch universities is worth mentioning. Furthermore, DK1 is a technical university. This again gives evidence of the subject-independence of the method since applied and technical sciences generally attract less citations than the natural sciences and most notably the life sciences.

### *Profiling excellence at the individual level*

Career evolution of individual scientists and changing team composition along with the typically small paper sets underlying the citation distribution at this level require extremely stable and robust solutions. In this section of the paper, we will analyse in how far the CSS-based method meets these requirements, and will give examples for its application to the level of individual scientists. We will show that the method can be applied using reference standards defined by any appropriate base-line distribution forming a superset of the publication set under study.

In addition, at the level of research teams and individual scientists bibliometric standard indicators can event more easily be distorted by the citation impact of papers that considerably differ from their expected value. In the previous section of this paper, we have the usefulness of the four performance classes that replaced the traditional mean-value based indicators at the level of national and institutional assessment of citation impact and the method proved to be insensitive to both subject-specific peculiarities and the particular choice of publication years (2007 vs. 2009) and citation windows (3 years vs. 5 years). The application to the micro level, that is, to the level of individual scientists and research teams, however, still remains a challenging task (cf. Wouters et al., 2013).

The selection of a proper benchmark is the first challenge. Traditionally we are faced with two typical tasks: firstly, de impact assessment in a multidisciplinary environment and, secondly, assessment in very specific and policy relevant research topics. As the first task implies a comparative analysis of individuals' research output in a given period, the underlying publication sets are from the statistical viewpoint, contrasting the cases discussed in the previous section, often too small to provide reliable results. Therefore the exercise has to be restricted to the group of more prolific authors. Yet, more prolific authors generally achieve higher visibility and impact, which, in turn, results in a somewhat "biased" baseline breaking the 70-21-9 rule. The reference standard thus depends on the selection criteria. In what follows, a typical example of such use case scenario will be described. The second case is somewhat easier if the research output used for the analysis is restricted to the topic under study. In this case, the topic, which is, otherwise, expected to attract significantly more citations than the discipline to which it belongs, should be used for defining the reference standard.

Deviations of the researchers' profile provide a multifaceted picture of their citation impact. A researcher's share in certain classes might be higher or lower than, or equal to the corresponding standard and his/her profile might thus follow the above-mentioned reference standard or be more or less polarised than the standard or more skewed towards poorly or highly cited papers, respectively. A researcher might have more highly cited papers than expected and at the same time less poorly cited papers than expected, but he/she might have more poorly cited papers then the reference standard. It should be mentioned that these cases also occur at higher levels of aggregations (cf. Denmark, Israel and Poland in Figure 1) but at the individual level they are more distinct and provide more pronounced information for a differentiated discussion.

An additional advantage is that the classes can, because of the high robustness of the distribution of papers across fields, be calculated at any level of field aggregation and that multiple field assignments do not hinder the calculation of the specific thresholds and the performance scores for individual authors. Furthermore, along with the selection criteria for the baseline model only the calculated threshold values are needed in a real world application of the method in an evaluative exercise without the underlying citation distributions.

Publications from authors with a Thomson Reuters Researcher-ID (cf. Heeffer et al., 2013) and at least 20 indexed publications were classified according to the field and year specific thresholds. This threshold has been chosen for reasons of statistical reliability. The data set was built on a total of 4.271 registered researchers active in all fields of science without any restriction on subject. When correlated with traditional normalised citation indicators it became clear that the distribution over classes is only partially correlated with these, most notably with the normalised journal based citation indicators. The CSS model thus provides essentially more information than the traditional bibliometric toolbox. The archetypes of the deviation of observed profiles from the baseline is shown in Figure 4.

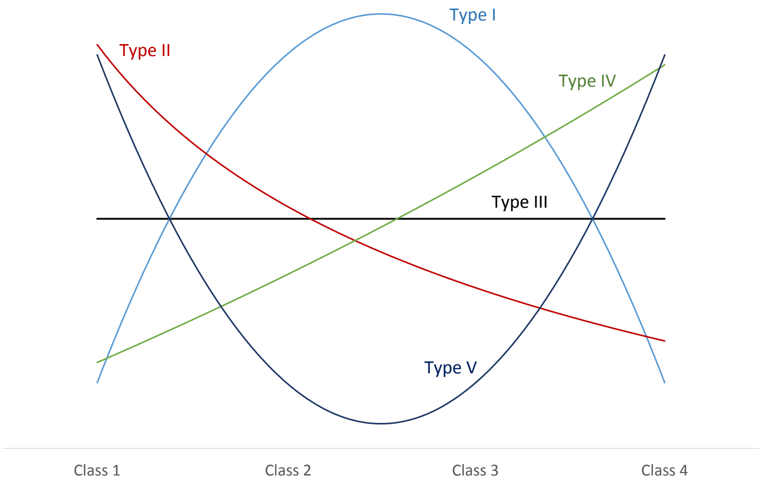


*Figure 4. The five different profiles according to their deviation from the reference standard (Type III is in line with the reference standard)*

Table 1 Average distribution of papers over performance classes organised by deviation type (benchmark values for the classes: 44.3%, 32.9%, 14.5% and 8.4%, respectively)

| Author # | Class 1 | Class 2 | Class 3 | Class 4 | Type |
|---|---|---|---|---|---|
| 1 | 31.8% | 27.3% | 36.4% | 4.5% | I |
| 2 | 39.3% | 60.7% | 0.0% | 0.0% | I |
| 3 | 32.5% | 45.0% | 22.5% | 0.0% | I |
| 4 | 40.0% | 24.4% | 33.3% | 2.2% | I |
| 5 | 67.5% | 28.6% | 1.3% | 2.6% | II |
| 6 | 58.3% | 32.5% | 7.3% | 2.0% | II |
| 7 | 74.2% | 16.1% | 9.7% | 0.0% | II |
| 8 | 72.0% | 20.0% | 8.0% | 0.0% | II |
| 9 | 68.2% | 27.3% | 0.0% | 4.5% | III |
| 10 | 33.3% | 52.4% | 9.5% | 4.8% | III |
| 11 | 61.9% | 38.1% | 0.0% | 0.0% | III |
| 12 | 33.3% | 28.9% | 20.0% | 17.8% | III |
| 13 | 21.4% | 21.4% | 14.3% | 42.9% | IV |
| 14 | 9.4% | 50.0% | 21.9% | 18.8% | IV |
| 15 | 17.9% | 14.3% | 28.6% | 39.3% | IV |
| 16 | 30.0% | 20.0% | 40.0% | 10.0% | IV |
| 17 | 56.8% | 22.4% | 11.2% | 9.6% | V |
| 18 | 62.5% | 16.1% | 12.5% | 8.9% | V |
| 19 | 47.6% | 9.5% | 9.5% | 33.3% | V |
| 20 | 50.0% | 10.7% | 14.3% | 25.0% | V |

Table 1 gives a (not random) sample of 20 authors taken from the above data set. In this example the total set of unique publications published by at least one of the authors from this set is used for benchmarking and its distribution over the four classes is taken as benchmark values. Therefore the reference set distinctly deviates from the overall baseline model with its 70–21–9 rule. In our example the reference was calculated based on all the publications of the selected authors. In this case, the set is a result from a within selection procedure but the reference set could also be defined prior to the start of this procedure, e.g., publications form a certain country or institute.

## 3. Discussion and conclusion

From its origins in support of library organization and information retrieval, scientometrics has developed into a scientific discipline that studies the characteristics and dynamics of scientific activity. The advent of large-scale publication databases coupled to the ever-increasing sophistication in indicator development, the advent and the development of advanced statistical, data-mining, text-mining and machine learning techniques have enabled the discipline to thrive. One driving force in this process is the need for quantitative mapping and assessment of scientific activity and its necessary prerequisite was the development of information technology that brought us the necessary data storage, availability of powerful computer equipment, worldwide communication and net-working and the widespread access to large databases. As a consequence, scientometrics has evolved from a sub-discipline of library and information science to an instrument for evaluation and benchmarking in support of science policy. This evolution, though, requires the development of indicators that allow to map and assess both institutional and individual profiles of scientific activity and visibility that neutralize the traditional pitfalls related to the toolbox of indicators that have traditionally been used to that purpose. Without such robust and valid methods and indicators, the use of scientometric data for science policy purposes remains prone to criticism and doubt.

In the present study we have described a novel indicator base that copes with some frequent pitfalls in evaluative scientometrics. We have compared national, institutional and individual impact profiles using typical examples. Further examples can be found in the literature referred to in the text. We also compared individual authors with a chosen reference standard and could detect five different paradigmatic profile types. Finally, we would like to stress that the reduction of this method to two instead of four performance classes would bring us back to system of traditional indicators.

The observations presented in this paper confirm the seamless integration of the CSS method into the standard toolset of scientometric research evaluation. The main idea was and remains, finally, to step away from the traditional linear thinking by depicting reality in a more differentiated way.

The analysis of the high end of scientific distributions is indeed one of the most difficult and challenging issues in evaluative scientometrics. This is, of course, not merely a mathematical issue as it is always difficult to draw a sharp borderline between “very good” and “outstanding” performance. Also the effect of outliers, i.e., of observations that might bias or even distort statistics, impressively shown by Waltman et al. (2012), is not typically a bibliometric issue. So-called censored data or data distorting extreme values of a distribution are known in several fields. In the proposed CSS-based method, the effect of outliers is limited since the influence of individual observations on the total is marginal, while the observations for the units under study are represented by classes instead of individual values.

Self-adjusting classes, such as those based on CSS, allow the definition of proper performance classes without any pre-set thresholds. This is certainly one of the main advantages of the proposed method. Another one is the seamless integration of measures of outstanding performance into the assessment tools of standard performance. The method of “implicit” subject fractionation can also be used in the context of other publication and citation indicators, whenever the issue of multiple subject assignment needs to be resolved.

Our studies have shown that a publication output at the meso-level suffices to provide a solid basis of interpretation and further statistical analysis. A further important property has become apparent, namely the method’s independence of the unit’s research profile. For instance, in small meso-level samples we have found two technical universities with more favourable citation profiles than that of medical universities or than their corresponding national reference standards. All those insights are judged relevant and important when engaging in evaluative scientometrics.

To conclude, the theoretical and empirical basis described in this paper underpins the validity and the robustness of the CSS method as a novel and relevant indicator base to measure, map and profile citation classes and their relationship to visibility and excellence of scientific output of countries, institutions, teams and individuals.

## References

Albarrán, P., & Ruiz-Castillo, J. (2011). References made and citations received by scientific articles. JASIST, 62(1), 40–49.

Glänzel, W. & Schubert, A. (1988), Characteristic Scores and Scales in assessing citation impact. *Journal of Information Science*, 14(2), 123–127.

Glänzel, W. (2007), Characteristic scores and scales. A bibliometric analysis of subject characteristics based on long-term citation observation. *Journal of Informetrics*, 1(1), 92–102.

Glänzel, W. (2013a), High-end performance or outlier? Evaluating the tail of scientometric distribution. *Scientometrics*, 97(1), 13–23

Glänzel, W. (2013b), *The Application of Citation-Based Performance Classes in Disciplinary and Multidisciplinary Assessment*. In: J. Gorraiz, E. Schiebel, Ch. Gumpenberger, M. Hörlesberger, H.F. Moed (eds), Proceedings of ISSI 2013 – The 14th International Conference on Scientometrics and Informetrics, Vienna, Austria, Vol. I, 109–122.

Glänzel, W., Thijs, Debackere, K. (2014), The application of citation-based performance classes to the disciplinary and multidisciplinary assessment in national comparison and institutional research assessment. *Scientometrics*, 101(2), 939–952.

Heeffer, S., Thijs, B., Glänzel, W. (2013). Are registered authors more productive? *ISSI Newsletter*, 9(2), 29–32.

Leydesdorff, L., Bornmann, L., Mutz, R., Opthof, T. (2011), Turning the tables on citation analysis one more time: principles for comparing sets of documents. *JASIST*, 62(7), 1370–1381.

Persson, O., Glänzel, W., Danell, R. (2004), Inflationary bibliometric values: The role of scientific collaboration and the need for relative indicators in evaluative studies. *Scientometrics*, 60(3), 421–432.

Thijs, B., Debackere, K., Glänzel, W. (2014). *Improved author profiling through the use of citation classes*. In: E. Noyons (Ed.), Proceedings of the STI Conference 2014, Leiden, 616–622.

Waltman, L., Calero-Medina, C., Kosten, J., Noyons, E.C.M., Tijssen, R.J.W., van Eck, N.J., van Leeuwen, Th.N., van Raan, A.F.J., Visser, M.S. & Wouters, P., (2012), *The Leiden Ranking 2011/2012: Data collection, indicators, and interpretation*. In: Eric Archambault, Yves Gingras, & Vincent Lariviere (Eds.), Proceedings of STI 2012 Montreal (17th International Conference on Science and Technology Indicators), Volume II, 791–802.

Wouters, P., Glänzel, W., Gläser, J., Rafols, I. (2013), The Dilemmas of Performance Indicators of Individual Researchers – An Urgent Debate in Bibliometrics. *ISSI Newsletter*, 9 (3), 48–53.