The value and accuracy of key figures in scientific evaluations

Jan Reedijk[1]
Leiden Institute of Chemistry, Leiden University, The Netherlands

Introduction

The use of bibliometric analyses in the evaluation of research performances of scientists and groups of scientists has been common practice in many countries and for a number of decades already. Such analyses were initially performed by just counting the number of publications in refereed journals; during the 1970s, counting of citations became possible and fashionable, and even averaged numbers of citations per published paper started to be used by purchasing the services of certain specialized institutions.

With the now common and user-friendly availability of the WoS (Web of Science), searches, analyses and evaluations are possible and 'easily' being carried out by the non-expert. In fact, with competing sources for data mining and analysis (sometimes cheaper, but less complete, such as Elsevier's Scopus and Google Scholar) now, everybody with access to the Internet can do amateur evaluations of individual (and groups of) scientists. It is often not realized that the consequences for scientists (evaluation, salary, grants, etc.) can be quite dramatic, especially when such analyses are incomplete, or erroneous.

In many countries and universities, and even international organizations such as the European Union in the FP7 (Seventh Framework Programme), salaries and grants, or grant renewals, are quite directly determined by bibliometric figures. Irrespective of whether this development is desirable, the accuracy and value of such figures need to be unquestionable. However, it appears that increasingly the use of such evaluation appears to be more a matter of the 'values' of the numbers, and not anymore of real 'value'. It should also be realized that the more simplified and metric-driven the evaluation of scientific work becomes, the more susceptible science will be to fabrication, tricks and even fraud.

What types of values are relevant, and how prone they are for database errors, misuse and abuse will be discussed below. Some of these issues I have addressed in recent papers [1–3]; these and others are the subject of this account. A key issue of course should be that if bibliometric figures are used in evaluations, their meaning should be clear, their accuracy should be high, and they should allow a fair comparison between scientists who are working in the same field and that are of the same (scientific) age.

Accuracy is a very important issue of course. Given the fact that references are still largely imported manually in the WoS, where typos in names or initials

[1]Email: reedijk@chem.leidenuniv.nl

can easily miss papers by authors, the accuracy of papers and citations is far less then desired. In addition, it is a known fact that authors when citing papers make typos, and, even worse, may copy and paste poorly from other papers, thereby incorrectly citing references. Moreover, it is well known that Thomson Reuters may, or may not, change names of institutions, when authors have used uncommon or old names of their institutions. This may result also in missing citations.

Parameters used in evaluations of scientists: the impact factor of journals

It is generally agreed that scientists should publish, and the more papers a scientist has (co-)authored in highly ranked and highly respected scientific journals, the higher the appreciation for the scientist will be. Therefore I first need to discuss the value of the most commonly used impact factor, the 2-year impact factor published on the WoS by Thomson Reuters, nowadays also called TRIF (Thomson Reuters Impact Factor) [4].

In many evaluations, each paper of a certain author is given a multiplication factor, which could be just the most recent TRIF value, or an integer, arbitrarily determined. For instance, journals with a TRIF below 3.00 obtain a multiplication factor of 1, journals with a TRIF higher than 2.99 and lower than 6 obtain a multiplication factor of 3, and so on. As TRIF factors change by definition every year (see below), it is important that the correct values are used for past years (they cannot be found easily for more than 5 years back). Of course, the biggest uncertainty here is by far not the inaccuracy in the TRIF, but rather the fact that individual papers in a journal can have dramatic differences in citations. So when a paper is published in a journal with a TRIF value of 7 for that year, this does not imply that all papers in that journal in that year would have the same numbers of citations, namely 7!

Literature on definitions and use of impact factors are plentiful. The impact factors can cover 1, 2, 5 or more years, they can be generated from Thomson Reuters, from Google Scholar or from the Elsevier Scopus databases. I just give the common one from TRIF here, and I refer to others for details and critical discussion [4].

Say for the year 2012, TRIF = C/P, where C = *citations in 2012 to papers appeared in the journal during 2010 and 2011 and* P = ***citable*** *publications in the journal during 2010 and 2011.*

To illustrate how risky the use of TRIF is when used for evaluations of a person, and also how meaningless TRIFs can be, I will discuss two recent examples of enormous jumps in TRIF lasting just 2 years (first up and, 2 years later, down again), each as the result of 'explosions' caused by a single article.

This jump happens when a super-hot (usually methodological or review) paper is cited extremely frequently. This high citation first may have an effect on the immediacy index, but subsequently, the TRIF jumps up and after that will go down again, just from the effect of that single paper. The two examples below from recent years clearly illustrate this effect; I already alluded briefly to one of these in 2012, even before the TRIFs for 2011 had been published [2]; TRIFs for 2012 appeared in June 2013 [5].

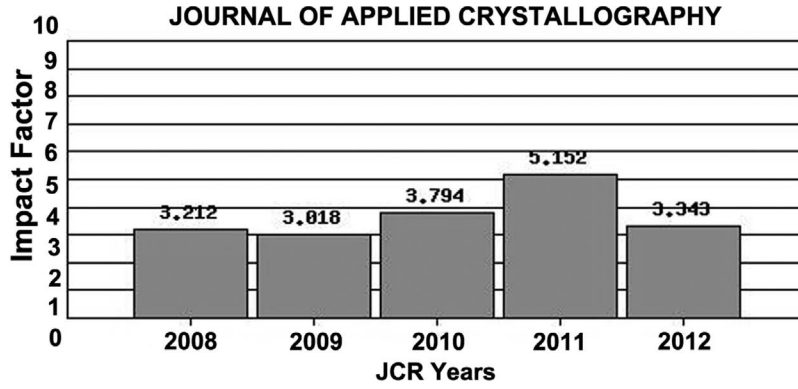
The first example deals with a methodological paper by Westrip [6], published in 2010 in the *Journal of Applied Crystallography*, a journal with some 200 papers per year. This single paper generated a huge number of citations already in 2010, but in particular in 2011, it received 712 citations, compared with 1086 citations made to all other 191 papers from 2010 in that journal. The consequence is a significant jump in TRIF, as seen in Figure 1.

An even more striking case is the paper by Sheldrick [7], a leading crystallographer. With his paper from 2008 in *Acta Crystallographica Section A*, a specialist journal, with approximately 50–70 papers each year, he generated a citation explosion bringing the TRIF in 2009 and 2010 up to 50 or higher, whereas before 2009 it was around 2, and, as was predicted beforehand, for the TRIF 2011 it returned to around 2 again (Figure 2). The paper was published in January 2008; the citations to it were: 3521 (in 2008), 4891 (in 2009), 6937 (in 2010), 8181 (in 2011) and 4816 (in 2012). It should be noted that the still high number of citations to this paper in 2012, by definition does not contribute to the TRIF 2011. From all the citations to this 2008 paper, only the citations obtained in 2009 and 2010 have an effect on the TRIF. That is, 11 828 citations to some 200 papers, and 95% of these are to this single paper!

It should be evident that the scores of these two single papers do not say anything at all about the quality (or citations) obtained by the other papers in that journal during these 2 years, although the jump in TRIF would suggest to many bureaucrats, administrators and even scientists that all papers in that journal were responsible for the increase in TRIF to 50! This misconception may also have a dramatic effect on the evaluation of scientists in the field, as I will show now.

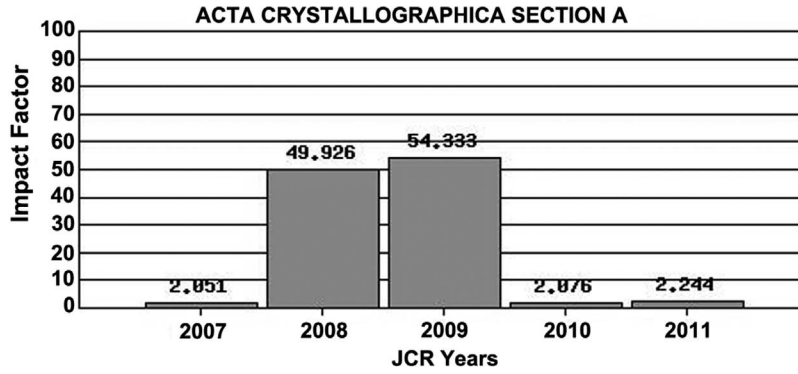
A hypothetical, but not unrealistic example to illustrate financial consequences: assume two authors in the same field, A and B, being equal in age, experience and quality. They each publish two papers in the above-mentioned *Acta Crystallographica Section A* and these two papers appeared within 1 month for each author. Coincidentally, author A has published the first paper in December 2008 and the second paper in January 2011; author B has published the first paper in January 2009, and the second in December 2010. Irrespective of the fact that

Figure 1



TRIFs for *Journal of Applied Crystallography* from 2008 to 2012
The temporary increase in 2011 is the result of the large number of citations to a single paper in 2010 [6].

Figure 2



TRIFs for *Acta Crystallographica Section A* from 2007 to 2011 [7].

whether these four papers are cited a lot, or perhaps not at all, the papers of author B would have been 'multiplied' by a factor of 50 in value; if she/he lives in a country where the salaries and grants are related to papers and high TRIF values, then B would have earned a fortune, whereas A would not have received this bonus!

These are just a few cases to emphasize that the TRIF value for a journal in a certain year does *not* imply that all papers in that journal over that year have the same quality. The TRIF of a journal is just a surrogate measure of the quality of its papers. It only shows their **average** citations. As early as 2005, *Nature* [8] wrote in an editorial: "Research assessment rests too heavily on the inflated status of the impact factor." They calculated that just 25% of the published articles in the period at hand contributed to 89% of the journal's 2005 impact factor [8]. Nevertheless, the process still goes on! And when the process continues, editors will find clever, but not always seen as fair, ways to optimize the TRIF, as discussed elsewhere [2,3]. Of course, we should not discourage our scientists from publishing in high-impact journals. We only should realize that for individual papers, a high-impact paper is not equivalent to the TRIF value of the journal.

At this point, I also want to mention another journal comparison parameter, SNIP (source normalized impact per paper), introduced by Moed [9]. SNIP measures the contextual citation impact of a journal, taking into account characteristics of its properly defined subject field, especially the frequency with which authors cite other papers in their reference lists. It covers in particular the rapidity of the maturing of the citation impact. Readers interested in this rapidly evolving discussion are referred to the original paper of Moed [9] and to later papers citing this work. Also, publishers' websites, such as Elsevier with SciVerse and Scopus, deal with recent developments in other parameters.

Parameters used in evaluations of scientists: the h (Hirsch)-index for (groups of) persons

A very recently and successfully introduced parameter is the h-index (h). The definition of h is as follows [10]: "*The number of papers (h) by the scientist that have received at least h citations in a given period.*"

Although initially set up for a single scientist, and meant for a whole career [10], the h-index can also equally well be used for groups of scientists, for a certain period, and even for a journal. Later, several derivatives of the standard h-index were also proposed [11–14]; they are not discussed in the present chapter. For an alternative to h-index, see a new suggestion by Bornmann [15] proposing the indicator $P_{top10\%}$ as a better substitute for the h-index. $P_{top10\%}$ is the number of publications which belong to the top 10% of most frequently cited publications. So a publication belongs to the top 10% of the most frequently cited only if it is cited more frequently than 90% of publications published in the same subject area and in the same year. Just like the h-index, $P_{top10\%}$ provides information about productivity and citation impact in a single number. So it just gives the number of publications for an author that are of major significance [15]. Whether this will supersede the h-index remains to be seen of course.

The (meanwhile) standard h-index appears to be very simple to calculate, one would think, but often it is not! The value of h can be used by research councils, for instance to allow submission of certain grant applications only for scientists with an h-value above a certain threshold number. The major caveats here are numerous. Let me just mention the following ones:

1. The standard h-index is age dependent; it can only increase over time; so older scientists generally have higher h-indices than younger colleagues in the same field.
2. The h-index is strongly research-field dependent; in some fields the number of citations per paper on average is much lower (as in mathematics) than in other fields (such as clinical medicine).
3. Citations to papers are necessarily all positive or relevant.
4. Scientists with a common name and common initials can have colleagues with the same name and initials, so their citations and papers are added up or mixed up. Even Thomson Reuters does so! Just to illustrate, the top 12 most-cited chemists worldwide, according to WoS on 1 May 2013 are all 'persons' with the names: Wang (J, Y and L), Kim (J), Zhang (L, J and Y), Liu (Y), Li (Y and J) and Chen (Y), i.e. all typical 'multiple' persons. In the current top 20, in fact only three 'single scientists' are present.
5. The h-index does not take into account the number of authors on any paper.
6. The h-index also does not consider the real numbers of citations.
7. The h-index is biased towards researchers that are active in writing review articles.

The use of the h-index and its many derivatives [14] has been critically discussed many times, and this will not be repeated here. Interested readers are referred to comments of Waltman, Marx and others [11,12,16] and many references cited there. To illustrate the large number of discussions: as many as 923 journal articles in the WoS by 1 May 2013 have 'h-index' in the title. The original Hirsch paper [10] had been cited over 1200 times by 1 May 2013, i.e. much more than any of Hirsch's research papers since 1960.

In addition to h values generated from the WoS (Thomson Reuters), Elsevier Scopus and Google Scholar also now allow the generation or calculation of h values, either directly at their website (Scopus), or via third-party software [17], or using a personal profile (Google Scholar). It appears that the values generated from Scopus or Google Scholar can differ significantly from those of Thomson Reuters, especially for older scientists, as Thomson Reuters has the most complete database of older papers. More importantly, problems do occur with authors who have over 1000 documents (Google Scholar), or with authors having common names and initials (all databases) so that proper filtering is required. It should be noted that from the database of Google Scholar, the h value can only be calculated for yourself and after generating an author profile. The use of a software package from Harzing [17] allows a quick, albeit less accurate, method to find h values. Unfortunately, such a search appears to be full of errors (more documents, and thus more, incorrect, citations in many cases). Moreover, for searches for names such as 'John Smith' or 'Peter Williams', it is almost impossible to perform quick analyses of the h-index using this software package [17], despite the rapid improvements of this site.

As a test case, I have checked my own data for the three databases, as per 1 May 2013. By using WoS, I arrive at a value of **79**. Elsevier Scopus makes it only **66**, and using Google Scholar after having generated a personal profile and removed manually non-research papers and chapters not belonging to me, this public value reaches **74** for me.

I have also performed the same analyses for 8 chemists, all having unique names and initials and working in my own fields of interest, but presented anonymously here. The quite large and striking variations in outcome are detailed in Table 1 below.

In the case of Google Scholar, too many 'papers' are found, as they include book chapters and patents, but often also concern just internal papers or local ones, or the hits even include all chapters of an edited book. For a more detailed comparison between WoS and Google Scholar, I refer to the Publish or Perish website [17].

The main general conclusion from the data in Table 1 is that quite large variations in h values are found, depending on the database used, for the same

Table 1
Differences in the h-index, calculated from different sources, as per 1 May 2013

Database		WoS		Scopus		Google/Harzing	
Name	Sub-field	Papers (n)	h	Papers	H	Documents	h
Author 1	Materials	587	83	553	75	886	89
Author 2	Organic	409	74	343	65	760	83
Author 3	Catalysis	356	55	377	54	527	58
Author 4	Inorganic	259	48	231	41	314	47
Author 5	Inorganic	224	33	180	29	251	33
Author 6	Organic	291	39	295	41	352	41
Author 7	Organic	676	88	584	81	888	91
Author 8	Organomet	208	47	198	31	297	43

WoS and Scopus data were directly taken from their websites. Google Scholar data were taken via the Publish or Perish software from Harzing [17].

scientist. Therefore the use of h values should not be recommended for use in evaluations, certainly not when the data generation is performed by non-specialists, or when the source is not indicated, and certainly not when the scientists have not been allowed to check the data before use.

Misuse and abuse

It is now increasingly common that metrics are used in evaluations of scientists, and that often only simple parameters are used. Therefore the temptations to try to influence these parameters will increase, and fabrication or engineering of data may be the result. So it can be predicted that the more simplified and metric-driven the evaluation of scientific work becomes, the more susceptible science will be to fabrication of data, tricks and even fraud.

To generate increases in the TRIF of their journal, editors can use methods where citations to the previous 2 years (i.e. those contributing to TRIF) are added in editorials, or they simply invite more reviews, which are known to be cited more in the first few years. Or they can encourage authors to 'not forget to cite this journal'. In case editors behave unethically in this respect, Thomson Reuters may temporarily remove (de-list or suppress) the journal from their annual listing in JCR (Journal Citation Reports); in fact, they have done so already in a quite few cases during the last decade [18]. A striking case is the *Journal of Gerontology*, where the departing editor in an editorial cites many papers from the previous 2 years [19]. According to the JCR website [20], the 2013 title-suppression list contained about 50 journals. Some journals, however, seem to challenge this behaviour of Thomson Reuters [21], as is also discussed in detail at The Scholarly Kitchen website [22].

Not really misuse or abuse, but certainly not correct is the following case. In the European Union FP7, the bureaucrats ask, or encourage, that applicants mention in their application, as part of their CV, the TRIF (of the journal for each of their papers). This should of course be the TRIF of the year when the paper was published. However, FP7 does not require such details, and as a result, applicants will most likely mention the most recent TRIF, even for papers that appeared a decade ago. Most journals have shown increasing TRIF values over the years, simply because the general trend in all fields is that the number of references per paper is increasing each year.

To increase the personal h value, certainly when the number is still low, say below 15 or 20, one could ask friends and colleagues to preferentially cite one's papers, or also increase self-citations. This is indeed also to be seen as unethical behaviour, but I am fairly sure it has been done and will still be done.

Concluding remarks

As shown above, scientific evaluations of scientists, groups of scientists, institutions and also journals are increasingly performed with the use of a single parameter, be it a so-called h-index for (groups of) scientists, or the TRIFs of the journals that the scientist(s) at hand used to publish their research results.

It was shown that the use of such parameters is full of risks, not only owing to often-significant inaccuracies and errors, but nowadays increasingly also because such parameters are subject to manipulation/fabrication. In fact, nowadays many organizations try to develop and use indices by counting things that can be counted, rather than considering factors that cannot be counted, but that are much more important. So it is to be expected that evaluations will probably become more and more a matter of values (numbers), and less so of value. For the time being, it is very important to be aware of these uses and consequences. For authors, I would therefore make the following recommendations.

1. If authors select a journal to submit a paper, they should not pay much attention to the (most recent) TRIF of the journal.
2. Authors should never change (the spelling of) their name (females) or initials or hyphenation (Hispanics) after a first publication has appeared; it may cost citations. Also authors should carefully check the spelling of their name in papers where they are a non-submitting co-author.
3. If scientists have to provide an h-index for their work at a certain date, they should always mention the measuring date and the database from which this value of h was generated.

Finally, I would also provide a few recommendations for science evaluators, selection committees, research councils and university boards.

1. Never use single parameters to evaluate scientists and applicants for jobs or grants. Realize that the data behind the derived parameters can be incorrect. I warned about this many years ago, as far back as 1998 [23], and my warnings were followed by those of many others [4,11,16,24–27] and I am far from being complete here. At this point, I should particularly mention also a recent initiative, received after writing the present chapter, coined the 'San Francisco Declaration on Research Assessment' (or DORA) from May 2013, signed by a large number of editors in the field of cell biology and with a petition to Thomson Reuters to change the journal impact factor significantly. For details and links, see their website regarding JCR numbers [28].
2. In case numerical data are desired to assist your scientific assessments, make sure you have professional assistance, and most importantly have the involved scientists themselves check such data, before they are used in evaluations. If one requires applicants to provide an h value, have the method of calculation (WoS, Scopus or Google Scholar) and the measuring date be explicitly mentioned.
3. Given the fact that citation data and parameters derived from such data can be manipulated and engineered, and that there are no reasons to assume that such engineering can be stopped, great care should be taken, also by the experts, in using numerical data for evaluations of individual scientists, departments and universities.

References

1. Reedijk, J. (2012) Citations and ethics. *Angewandte Chemie International Edition* **51**, 828–830
2. Moed, H.F., Colledge, L., Reedijk, J., Moya-Anagon, F., Guerrero-Bote, V., Plume, A. and Amin, M. (2012) Citation-based metrics are appropriate tools in journal assessment provided that they are accurate and used in an informed way. *Scientometrics* **92**, 367–376
3. Reedijk, J. and Moed, H.F. (2008) Is the impact of journal impact factors decreasing? *Journal of Documentation* **64**, 183–192
4. VanClay, J. (2012) Impact factor: outdated artefact or stepping stone to journal certification. *Scientometrics* **92**, 211–238
5. Thomson Reuters (2013) Release of Journal Citation Reports, 2012 Citation Data In Journal Citation Reports®. http://admin-apps.webofknowledge.com/JCR/static_html/notices/notices.htm
6. Westrip, S.P. (2010) publCIF: software for editing, validating and formatting crystallographic information files. *Journal of Applied Crystallography* **43**, 920–925
7. Sheldrick, G.M. (2008) A short history of SHELX. *Acta Crystallographica Section A* **64**, 112–122
8. Editorial (2005) Not-so-deep impact. *Nature* **435**, 1003–1004
9. Moed, H.F. (2010) Measuring contextual citation impact of scientific journals (SNIP). *Journal of Informetrics* **4**, 265–277
10. Hirsch, J.E. (2005) An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 16569–16572
11. Abramo, G., D'Angelo, C.A. and Viel, F. (2010) A robust benchmark for the h and g indices. *Journal of the Association for Information Science and Technology* **61**, 1275–1280
12. Bornmann, L. and Marx, W. (2013) How good is research really. *EMBO Report* **14**, 226–230
13. Molinie, A. and Bodenhausen, G. (2010) Bibliometrics as weapons of mass citation. *Chimia* **64**, 78–89
14. Bornmann, L., Mutz, R., Hug, S.E. and Daniel, H.D. (2011) A multilevel meta-analysis of studies reporting correlations between the h index and 37 different h index variants. *Journal of Informetrics* **5**, 346–359
15. Bornmann, L. (2013) A better alternative to the H index. *Journal of Informetrics* **7**, 100
16. Waltman, L. and van Eck, N.J. (2012) The inconsistency of the h index. *Journal of the Association for Information Science and Technology* **63**, 406–415
17. Harzing, A.W. (2007) Publish or Perish. http://www.harzing.com/pop.htm
18. Agrawal, A.A. (2005) Corruption of journal impact factors. *Trends in Ecology & Evolution* **20**, 157
19. Morley, J.E. (2004) Flying through 5 years. *Journals of Gerontology. Series A, Biological Sciences and Medical Sciences* **59**, 1270–1276
20. JCR (2013) Suppression list Journals. *Journal Citation Reports®*. http://thomsonreuters.com/press-releases/062013/2013-journal-citation-reports
21. van der Wall, E.E. (2012) The NHJ 2012 in retrospect: which articles are cited most. *Netherlands Heart Journal* **20**, 481–482
22. Davies, P. (2013) *Netherlands Heart Journal* editor delivers Dutch citation treat. http://scholarlykitchen.sspnet.org/2013/01/30/netherlands-heart-journal-editor-delivers-dutch-citation-treat/
23. Reedijk, J. (1998) Sense and nonsense of science cititation analyses: comments on the monopoly position of ISI and citation inaccuracies. Risks of possible misuse and biased citation and impact data. *New Journal of Chemistry* **22**, 767–770
24. Bornmann, L., Mutz, R. and Daniel, H.D. (2008) Are there better indices for evaluation purposes than the h index? A comparison of nine different variants of the h index using data from biomedicine. *Journal of the Association for Information Science and Technology* **59**, 830–837
25. Bornmann, L., Mutz, R. and Daniel, H.D. (2009) Do we need the h index and its variants in addition to standard bibliometric measures? *Journal of the Association for Information Science and Technology* **60**, 1286–1289
26. Ernst, R.R. (2010) The follies of citation indices and academic ranking lists. A brief commentary to 'Bibliometrics as Weapons of Mass Citation'. *Chimia (Aarau)* **64**, 90
27. Kotov, N.A. (2010) Fraud, the h-index, and Pasternak. *ACS Nano* **4**, 585–586
28. San Francisco Declaration on Research Assessment (draft). http://tagteam.harvard.edu/hub_feeds/119/feed_items/165388