Contents lists available at SciVerse ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

Calibration against a reference set: A quantitative approach to assessment of the methods of assessment of scientific output

Marek Kosmulski

Lublin University of Technology, Lublin, Poland

ARTICLE INFO

Article history:
Received 14 November 2011
Received in revised form 19 February 2012
Accepted 20 March 2012

Keywords:
Citation analysis
Scientific output
Hirsch-type index

ABSTRACT

A set of authors whose scientific output can be unequivocally ranged from the highest to the lowest is used to assess the methods of assessment of scientific output. The rank-rank correlation coefficient between the known order in the calibration set and the order produced by certain method of assessment is a quantitative measure of the quality of that method. A common-sense-based reference may play a positive role in the communication between the enthusiasts and antagonists of bibliometric indices.

1. Introduction

The opinions about the quality of different methods of assessment of scientific output (based on peer review, number of citations, etc.) are subjective, and sometimes extreme (very positive and very negative) opinions are expressed by different authors about the same method. Extreme opinions are often accompanied by examples. The problem is that practically each method of assessment can be easily "substantiated" or "challenged" by an especially designed example. The divergence of opinions about particular methods of assessment of scientific output is in line with narcissistic side of the human nature: we prefer methods in which our own scientific output receives a high score. The scores obtained by means of different (reasonable) methods are to some degree correlated, but selection of this or another method may favor or disfavor individuals.

Also new methods of assessment of scientific output are introduced in the literature by means of examples, which are supposed to prove the superiority of a newly introduced method over the old methods. The examples are arbitrarily selected and they differ from one paper to another, thus their significance and persuasive power are limited.

A new approach to the "assessment of assessment" of scientific output of individuals is discussed in this paper. Let us define a reference set of authors, whose scientific output can be easily and unequivocally ranged from the highest to the lowest. Certainly many different reference sets can be defined. Let us set the following limitations:

1. All authors in the set represent the same branch of science. This limitation reflects the difference in publication and citation cultures between different branches of science.
2. All authors in the set are in similar biological and scientific age.

The above conditions imply relative character of the criteria of assessment: the method suitable to assess astronomers, who started their scientific careers in the 1960s may be completely unsuitable to assess mathematicians, who started their careers in the 1990s. The choice of the reference set is a difficult compromise between large branches of science (e.g., chemistry) and wide ranges of age (e.g., 20 years) on the one hand, and narrow sub-branches and narrow ranges of age (e.g.

E-mail address: mkosmuls@abo.fi

1751-1577/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.joi.2012.03.004

physical chemists who published their first papers in 1978) on the other. A wide branch/range has a disadvantage of the difference in publication and citation cultures in its sub-branches (e.g. organic vs. analytical chemistry) and of the effect of authors' age on the bibliometric indicators, but it has an advantage of universality of the results. The choice of the reference set is also a difficult compromise between the number of authors in the set (a larger set is more sensitive to the difference between particular methods) and the obviousness of the order within the set.

The present approach is to some degree related to the studies of the correlations between the scores obtained in peer judgment on the one hand and bibliometric indicators on the other. Namely, the ranking based upon peer judgment is usually considered at the "correct" one. Such studies have been carried out at various aggregation levels ranging from single article (e.g., Allen, Jones, Dolby, Lynn, & Walport, 2009) to institutions (e.g., Van Raan, 2006), and several studies were carried out at aggregation level author (e.g., Bornmann & Daniel, 2005). The ranking lists of authors based on peer judgment are not particularly useful as calibration sets (in the sense discussed above) for the following reasons:

- The peer-judgment of authors serves certain purpose, which is usually different from assessment of overall scientific output.
- Usually the peer-judgment results in a few-level score, e.g., approved/rejected (Bornmann & Daniel, 2005) or 3–4–5 (Van Raan, 2006). This results in a ranking list with many tied ranks, which is not very sensitive.
- Usually the peer-judgment of authors refers to a group of volunteers (e.g. applicants for a grant or for a position), thus the group is not representative.

Therefore an especially created reference set seems to be a more suitable tool than existing peer-review-based ranking lists of scientists.

The order (from the highest to the lowest output) in a properly selected set can be established without invocation of specific bibliometric indicators, but the distances between particular authors cannot. Therefore, the rank-rank (rather than Pearson-type) correlation coefficient between the order in the calibration set and the order produced by certain method of assessment is used as a quantitative measure of the quality of that method. In the present paper the principle of the "assessment of assessment" is presented using 4 arbitrarily selected reference sets composed of four chemists each, and one larger set composed of all 16 chemists. The strategy of selection of a proper reference set is a problem by itself, and it requires deeper studies. Probably more powerful reference sets than those used in the present study can be constructed.

This present approach is different from the usual approach, in which the mutual correlations between the results produced by particular methods are studied. Bornmann, Mutz, Hug, and Daniel (2011) presented an overview of studies reporting correlations of various indices with h and performed meta-analysis of those studies. Most h-type indices are highly correlated with the original h-index. In the studies of correlations, none of the indices is considered as superior to the others.

2. Reference sets

The following reference set of authors was used in the present study.

N: a recent Nobel laureate in chemistry.

F: Polish professor of chemistry, Ph.D., Dr.Sci. (habilitation is a higher doctorate in Poland and in a few other European countries), a recent laureate of the prize of Polish Science Foundation (called "Polish Nobel prize"). No more than four prizes a year are granted by PSF including no more than one in: life sciences, natural sciences, humanities, and applied sciences.

P: Polish professor of chemistry, Ph.D., Dr.Sci. Has never received any prize comparable with those received by N or F.

D: Polish chemist. Has never received any prize comparable with those received by N or F. Has never received a Dr.Sci. degree or a position of university professor.

Each of four calibration sets discussed below is composed of one representative of each category (N, F, P, D). The four scientists in each set are in similar biological age. The sixteen scientists in all sets are also in similar biological age, but the age window is broader than in each individual set.

The order: N > F > P > D in terms of scientific output is obvious. Less obvious is if the present sets are representative for other sets composed of N, F, P and D. The individuals selected for this study have rather unique combinations of the 1st and last names so that their automatically created outputs needed only small corrections (if any). Thus a tedious analysis of sets of homonymous individuals, multiple spellings of the same name, etc. could be avoided. The number of sets is limited by the small number of N and F on the one hand, and by limited availability of the birthday information about non-prominent scientists on the other.

3. Results and discussion

The scientific outputs of the individuals described above were assessed using the following standard indicators, automatically displayed in the WoS database (accessed on February 18, 2012):

- Number of publications.
- Number of citations.

Table 1
Assessment of 18 indicators by means of a reference set 1 of authors (born 1940–1945).

Author	N1	F1	P1	D1	Score
Date of 1st paper	1969	1971	1969	1977	n/a
# papers	103	330	180	6	0.4
# citations	5596	6621	2448	220	0.8
Citations/article	54.33	20.06	13.6	36.67	0.4
h	21	45	27	2	0.4
m	0.49	1.1	0.63	0.06	0.4
h(2)	8	10	7	1	0.8
h(3)	5	4	4	1	0.95
Most cited paper	3345	296	112	216	0.8
# papers with >100 citations	6	9	1	1	0.74
g	74	66	40	6 (14)[a]	1
Jin's A	249.1	81.18	49.3	109	0.4
Jin's R	72.33	60.44	36.48	14.76	1
π	48.7	20.59	8.37	2.18	1
t	35	61	37	3	0.4
1st author h	11	20	20	2	0.32
NSP	35	91	32	1	0.8
% SP	33.98	27.58	17.78	16.67	1
International recognition	28	25	20	7	1
#1st ranks	9	8.5	0.5	0	
Rank in # 1st ranks	1	2	3	4	1

[a] The number in parentheses was obtained by addition of dummy papers with 0 citations each

- Number of citations per paper.
- h-index.

The following well-known modifications of the h-index were studied:

- m-index (the h-index divided by the career length in years) (Hirsch, 2005).
- g-index (the maximum number, for which the top g papers have together at least g^2 citations) (Egghe, 2006).
- A-index (the average number of citations in the h-core) (Jin, Liang, Rousseau, & Egghe, 2007).
- R-index (square root of the sum of the numbers of citations of papers in the h-core) (Jin et al., 2007).
- Pi-index (0.01 of the number of citations of "elite papers" defined as the top square root of the total number of papers) (Vinkler, 2010).
- t-index (the maximum number, for which the geometric average of the number of citation of the top t papers is at least t) (Tol, 2009).
- 1st author h (h index calculated for papers, in which the assessed scientist is the 1st author) (Opthof & Wilde, 2009).
- h(2)-index (the maximum number, for which the h(2)th most-cited paper has at least $[h(2)]^2$ citations) (Kosmulski, 2006).
- h(3)-index (the maximum number, for which the h(3)th most-cited paper has at least $[h(3)]^3$ citations) (Kosmulski, 2006).

The following indicators were studied:

- the number of citations of the most-cited paper.
- the number of papers cited >100 times each.

These indicators may be considered as a number of citations in a set of "elite papers", which is limited to one top paper, and the number of "elite papers" defined by an arbitrarily selected number of citations, respectively. The later approach is especially popular in Poland.

The following indices recently designed by the present author were studied:

- Hirsch-type index of international recognition (in the list of countries arranged by the number of papers citing certain scientist, the highest rank h, for which at least h papers from certain country cite that scientist) (Kosmulski, 2010).
- Number of significant papers (SP: a paper which has more citations than the number of references in that paper) (Kosmulski, 2011).

Moreover an additional indicator: % of significant papers was defined as the ratio NSP: number of papers published. The numerical values of particular indicators and the rank-rank correlation coefficient between the assumed and obtained order (score) are summarized in Tables 1–4. The list of indicators in Tables 1–4 was arbitrarily selected, and addition of rows to Tables 1–4 would not pose any theoretical of technical problem or change the existing rows.

Fourteen of the 18 indices have an extensive character, and they never decrease in the course of the scientific career. Thus they favor older over younger scientists. One indicator (m) is normalized to the career length, and it may decrease in

Table 2
Assessment of 18 indicators by means of a reference set 2 of authors (born 1943–1945).

Author	N2	F2	P2	D2	Score
Date of 1st paper	1965	1972	1971	1975	
# papers	254	213	283	31	0.4
# citations	11,194	4943	3139	306	1
Citations/article	44.07	23.21	11.29	9.87	1
h	49	40	29	10	1
m	1.04	1	0.71	0.27	1
h(2)	11	9	7	4	1
h(3)	6	5	4	3	1
Most cited paper	2083	175	92	49	1
# papers with >100 citations	23	7	0	0	0.95
g	100	60	39	16	1
Jin's A	168.92	74.3	45.34	25	1
Jin's R	90.98	54.51	36.26	15.33	1
π	58.39	16.35	9.09	1.78	1
t	76	56	38	15	1
1st author h	25	31	17	6	0.8
NSP	97	58	18	6	1
% SP	38.19	27.23	6.36	19.35	0.8
International recognition	34	28	16	6	1
#1st ranks	16	1	1	0	
Rank in # 1st ranks	1	2.5	2.5	4	0.95

the course of the scientific career. Three other indicators (citations per paper, % SP, and to some degree A) are normalized to the number of papers, and they punish over-production of low-impact papers.

Match in terms of both biological and scientific age was attempted within each set, but the success in this respect was limited, and the ranges of the scientific age are up to 10 years in particular sets. Generally those who sooner published their 1st papers were more successful in terms of bibliometric indicators, with a few exceptions.

Tables 1–4 ignore the indicators which involve corrections for self-citations and for multiple authorship. In the t-index, the role of the author is acknowledged rather than the number of co-authors. In the present reference set, the effect of self-citations and multiple authorship on the rank-rank correlations was limited. This is due to large gaps between particular authors in terms of the bibliometric indices considered in Tables 1–4, and to similar numbers of co-authors in the highly cited publications in the present reference set.

All indicators but one produced a positive rank-rank correlation with the assumed order in all sets. The expected value for two unrelated quantities equals to zero, thus any positive correlation coefficient can be considered as a "success". The only negative correlation was observed in Table 4 for the % SP-index, which was defined in this paper. The % SP-index was especially introduced to demonstrate that a reasonably looking bibliometric index can be negatively correlated with the assumed order. On the other hand, the number of perfect correlations with the assumed order ranges from 5 in Table 3 to 14 (out of 19) in Tables 2 and 4.

Table 3
Assessment of 18 indicators by means of a reference set 3 of authors (born 1945–1947).

author	N3	F3	P3	D3	Score
Date of 1st paper	1968	1972	1975	1975	
# papers	689	105	96	35	1
# citations	33,282	685	953	161	0.8
Citations/article	48.3	6.52	9.93	4.6	0.8
h	94	14	17	7	0.8
m	2.14	0.35	0.46	0.19	0.8
h(2)	16	5	5	3	0.95
h(3)	6	3	3	2	0.95
Most cited paper	904	79	68	31	1
# papers with >100 citations	87	0	0	0	0.77
g	149	22	27	11	0.8
Jin's A	188.5	28.64	37.24	14.57	0.8
Jin's R	133.11	20.02	25.16	10.1	0.8
π	87.56	3.25	4.53	0.95	0.8
t	134	20	25	9	0.8
1st author h	53	8	9	1	0.8
NSP	330	19	12	0	1
% SP	47.9	18.1	12.5	0	1
International recognition	29	12	11	6	1
#1st ranks	18	0	0	0	
Rank in # 1st ranks	1	3	3	3	0.77

Table 4
Assessment of 18 indicators by means of a reference set 4 of authors (born 1946–1947).

author	N4	F4	P4	D4	Score
Date of 1st paper	1970	1969	1979	1975	
# papers	554	143	87	3	1
# citations	26,420	7546	780	62	1
Citations/article	47.69	52.77	8.97	20.67	0.6
h	92	48	16	3	1
m	2.19	1.12	0.48	0.08	1
h(2)	14	11	5	2	1
h(3)	6	5	3	2	1
Most cited paper	609	759	39	41	0.6
# papers with >100 citations	80	20	0	0	0.95
g	132	83	21	3 (7)[a]	1
Jin's A	156.52	122.08	24.88	20.66	1
Jin's R	120.01	76.55	19.95	7.87	1
π	60.49	30.83	2.75	0.62	1
t	123	70	20	3	1
1st author h	22	15	13	2	1
NSP	250	58	8	2	1
% SP	45.13	40.56	9.2	66.67	-0.2
International recognition	35	26	11	4	1
#1st ranks	15	2	0	1	
Rank in # 1st ranks	1	2	4	3	0.8

[a] The number in parentheses was obtained by addition of dummy papers with 0 citations each.

Only in one set (Table 3), N was the winner in terms of all bibliometric indicators. In other sets, F, P and/or even D were winners in one or more bibliometric indicators.

The scores obtained in a set of all 16 scientists and in 4 smaller sets (Tables 1–4) are summarized in Table 5. In the calculation of the score for a reference set of 16 chemists, all four N were considered as tied rank 1–4, all four F as tied rank 5–8, etc.

Three different strategies of the data handling are presented in Table 5. The 3rd column is a sum of the scores from Tables 1–4. Large sum indicates a successful bibliometric index. The 4th column is a number of perfect correlations obtained in Tables 1–4. Certainly many perfect correlations indicate a successful bibliometric index. The disadvantage of this method is low precision (only five values are possible) and many tied ranks. The 5th column is a number of very low correlations obtained in Tables 1–4. A correlation coefficient of <0.5 was arbitrarily selected as an indicator of a "failure". Certainly a few failures indicate a successful bibliometric index. The two later criteria can be combined. A large difference between the number of full correlations and the number of failures indicates a successful bibliometric index.

The examples in Tables 1–5 demonstrate that even an apparently obvious and unequivocal ranking of scientists is not fully reproduced by standard bibliometric indicators. Only the index of international recognition produced perfect matches with the assumed order in all 4 reference sets, but not in the set of 16. It should be emphasized that because of tied ranks in the reference set of 16, the correlation coefficient of 1 with any bibliometric indicator is very unlikely.

Table 5
Assessment of 18 indicators by means of a reference set of 16 authors and summary of results from Tables 1–4.

	Overall	Tables 1–4		
	Overall score	Sum of scores	# scores = 1	# scores <0.5
International recognition	0.94	4	4	0
π	0.93	3.8	3	0
g	0.92	3.8	3	0
NSP	0.92	3.8	3	0
Jin's R	0.92	3.8	3	0
h(3)	0.90	3.9	2	0
Citations	0.9	3.6	2	0
h(2)	0.9	3.75	2	0
t	0.84	3.2	2	1
h	0.83	3.2	2	1
m	0.81	3.2	2	1
Most cited paper	0.81	3.4	2	0
Jin's A	0.80	3.2	2	1
# papers with >100 citations	0.78	3.41	0	0
# papers	0.77	2.8	2	2
1st author h	0.76	2.92	1	1
Citations/article	0.63	2.8	1	1
% SP	0.55	2.6	2	1
# 1st ranks	0.47	3.52	1	0

The overall assessment and all 3 methods based on the correlations observed in 4 smaller sets produce similar but not identical ranking of bibliometric indicators. The index of international recognition produced the highest scores in columns 2–4 and the tied highest score in column 5. The π, g, NSP, and R produced almost identical scores (and only marginally lower than the index of international recognition) in all columns in spite of being based upon completely different principles.

The famous h index produced rather moderate scores in all columns. The number of publications, 1st author h, and the number of citations per article produced overall correlation coefficients below 0.8, sums of correlation coefficients from Tables 1–4 below 3, and numbers of full correlations (score 1 in Tables 1–4) equal to the numbers of failures in sets 1–4, and they seem to be the least successful bibliometric indicators in terms of reproduction of the assumed order in the analyzed reference sets. The failure of the number of publications is due to under-appreciation of high-impact articles (N1) and over-appreciation of high production of low-impact articles (P2). The failure of the 1st author h is due to the tendency in leaders of Polish scientific institutions to put their own names as the 1st author, while in many other countries the group leader is the last author. The failure of the number of citations per article is due to undue appreciation of authors of one high-impact paper, and of very few other papers (D1).

The present results do not imply that the index of international recognition, π, g, NSP, and R (rank 1–5 in the 2nd and 4th column of Table 5, tied rank 1 in the last column, and rank 1–6 in the 3rd column of Table 5) are superior to other indicators studied in this paper. Probably another reference set of 16 mature chemists can be easily made up, for which h- and m-indices produce a perfect correlation with the expected order, and g, R and π do not. On the other hand the present reference set clearly indicates that the %-SP and the number of publications are rather useless as tools of assessment.

4. Further research

The advantage of the present method is that the reference set was composed using only common sense, e.g., without direct reference to numerical values of bibliometric indicators. Such an approach is of special importance in the communication between the enthusiasts of bibliometric indices and their antagonists. The present results confirm the usefulness of standard bibliometric indicators and of a few less-well-known bibliometric indicators for assessment of mature chemists. This result does not imply the usefulness of the same bibliometric indicators for assessment of other age groups or of representatives of other branches of science.

The arbitrariness is a clear weakness of the reference sets 1–4. Much larger reference sets can hardly be set up without having lost the obviousness of order in the reference set. Use of multiple reference sets rather than one large set is an attractive alternative to a large set. The correlation coefficients (scores) calculated for particular reference sets can be further processed to obtain the overall score, also using strategies other than those presented in Table 5.

An analogous approach can be considered for other aggregation levels than single author (e.g., institution, journal) as suggested by one of the referees of this paper.

References

Allen, L., Jones, C., Dolby, K., Lynn, D., & Walport, M. (2009). Looking for landmarks: The role of expert review and bibliometric analysis in evaluating scientific publication output. *PlosOne*, *4*(6), e5910.

Bornmann, L., & Daniel, H. D. (2005). Does the h-index for ranking of scientists really work? *Scientometrics*, *65*(3), 391–392.

Bornmann, L., Mutz, R., Hug, S. E., & Daniel, H. D. (2011). A meta-analysis of studies reporting correlations between the h index and 37 different h index variants. *Journal of Informetrics*, *5*(3), 346–359.

Egghe, L. (2006). Theory and practise of the g-index. *Scientometrics*, *69*(1), 131–152.

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.

Jin, B., Liang, L., Rousseau, R., & Egghe, L. (2007). The R- and AR indices: Complementing the h-index. *Chinese Science Bulletin*, *52*(6), 855–863.

Kosmulski, M. (2006). A new Hirsch-type index saves time and works equally well as the original h-index. *ISSI Newsletter #7* (pp. 4–6).

Kosmulski, M. (2010). Hirsch-type index of international recognition. *Journal of Informetrics*, *4*(3), 351–357.

Kosmulski, M. (2011). Successful papers: A new idea in evaluation of scientific output. *Journal of Informetrics*, *5*(3), 481–485.

Opthof, T., & Wilde, A. A. (2009). The Hirsch-index: A simple, new tool for the assessment of scientific output of individual scientists. *Netherlands Heart Journal*, *17*(4), 145–154.

Tol, R. S. (2009). The h-index and its alternatives: An application to the 100 most prolific economists. *Scientometrics*, *80*(2), 317–324.

Van Raan, A. F. (2006). Comparison of the Hirsch-index with standard bibliometric indicators and with peer judgment for 147 chemistry research groups. *Scientometrics*, *67*(3), 491–502.

Vinkler, P. (2010). The π_v-index: A new indicator to characterize the impact of journals. *Scientometrics*, *82*, 461–475.