Scientometrics (2011) 88:1011–1016
DOI 10.1007/s11192-011-0424-8

A comment to the paper by Waltman et al., Scientometrics, 87, 467–481, 2011

Tobias Opthof · Loet Leydesdorff

Received: 11 May 2011 / Published online: 17 June 2011

Abstract In reaction to a previous critique (Opthof and Leydesdorff, J Informetr 4(3):423–430, 2010), the Center for Science and Technology Studies (CWTS) in Leiden proposed to change their old "crown" indicator in citation analysis into a new one. Waltman (Scientometrics 87:467–481, 2011a) argue that this change does not affect rankings at various *aggregated* levels. However, CWTS data is not publicly available for testing and criticism. Therefore, we comment by using previously published data of Van Raan (Scientometrics 67(3):491–502, 2006) to address the pivotal issue of how the results of citation analysis correlate with the results of peer review. A quality parameter based on peer review was neither significantly correlated with the two parameters developed by the CWTS in the past citations per paper/mean journal citation score (CPP/JCSm) or CPP/FCSm (citations per paper/mean field citation score) nor with the more recently proposed *h*-index (Hirsch, Proc Natl Acad Sci USA 102(46):16569–16572, 2005). Given the high correlations between the old and new "crown" indicators, one can expect that the lack of correlation with the peer-review based quality indicator applies equally to the newly developed ones.

Keywords Citation · Indicator · *h*-index · Quality · Excellence · Selection

We react on a study by Waltman et al. (2011a), entitled "Towards a new crown indicator: An empirical analysis." The authors go at great length to show that a change in the

T. Opthof
Experimental Cardiology Group, Heart Failure Research Center,
Academic Medical Center AMC, Meibergdreef 9, 1105 AZ Amsterdam, The Netherlands
e-mail: t.opthof@inter.nl.net

T. Opthof
Department of Medical Physiology, University Medical Center, Utrecht, The Netherlands

L. Leydesdorff (✉)
Amsterdam School of Communication Research (ASCoR), University of Amsterdam,
Kloveniersburgwal 48, 1012 CX Amsterdam, The Netherlands
e-mail: loet@leydesdorff.net

normalization—in reaction to our previous critique of the Leiden "crown" indicators (Opthof and Leydesdorff 2010)—did not significantly affect the rankings at various *aggregated* levels. Since the Center for Science and Technology Studies (CWTS)-data under discussion were not publicly available,[1] let us use a previous occasion at which Van Raan (2006) revealed some of the micro-data underlying the evaluations in the case of 147 research groups in chemistry. The defense at that time was triggered by the introduction of the *h*-index by Hirsch (2005). How did the Leiden "crown" indicators work in comparison to the *h*-index? Unlike the citation indicators, the *h*-index is sensitive to the number of publications for which citation rates are compared. Decomposition of aggregated data allows for distinguishing mechanisms; for example, variance "within groups" versus "between groups."

Since Narin (1976) suggested the use of bibliometrics for evaluative purposes, semi-industrial centers have sprung up either connected to academia (such as in Budapest, Leiden, Leuven, Beijing, Shanghai, etc.) or as independent commercial enterprises (e.g., Science-Metrix in Montreal). Two major companies (Thomson Reuters and Elsevier) are also active in this market. In other words, citation analysis has become an industry. Intellectual property of the data and the results of the analysis has become a major asset in this (quasi-)industry. Although contractors sometimes state that the results are freely available for the users, the licenses of the data (the *Science Citation Index*) often do not permit to publish results freely so that the scientists under study would be able to control these evaluations themselves (cf. Opthof and Leydesdorff 2010). This practice of secrecy tends to shield the evaluation against the criticism that has been voiced against the use of citation analysis for evaluative purposes (Leydesdorff 2008; MacRoberts and MacRoberts 1987, 1996, 2010).

The invention of the *h*-index as a new statistics in 2005 (Hirsch 2005), however, challenged the leading researcher of CWTS (Van Raan 2006) to test whether this new indicator correlated with the "crown" indicators of scientometric evaluation in use by CWTS: citations per paper/field citation score (CPP/FCSm) and CPP/JCSm (Schubert and Braun 1986; Vinkler 1986; Moed et al. 1995). These latter indicators have extensively been used for such purposes as the Leiden Rankings of universities, research evaluation at the institutional level, and science policy advice at national and international (e.g., EU) levels (e.g., Moed 2005). Vinkler (1996) considered this indicator—which he indicated with RW—as the most appropriate one for the evaluation.

The CWTS study (VSNU 2002) was based on more than 18,000 publications of 147 research groups in chemistry and chemical engineering in the Netherlands for the years 1991–1998. A subset of this data was secondarily analyzed by Van Raan (2006). In addition to the citation indicators, the research groups under study were peer reviewed on their quality on a five-point scale. All fields within chemistry were covered by this set of university groups. The author notes that the various specialties exhibit different citation characteristics and that therefore field-normalization would be essential (cf. Leydesdorff and Opthof 2010, 2011). CPP/FCSm normalizes CPP for the mean FCSm where a "field" is defined as a set of journals sharing a field-code of the ISI Subject Categories. Analogously CPP/JCSm normalizes for the mean citation scores of individual journals (Schubert and Braun 1986; Vinkler 1986; Waltman et al. 2011b).

[1] One of us recently (Jan. 20, 2011) received access to this data in response to a request of the Dean of the Academic Medical Center of the University of Amsterdam. This communication was first submitted before that date.

Table 1 Example of the results of the bibliometric analysis for the chemistry groups

Research group	P	C	CPP	JCSm	FCSm	CPP/JCSm	CPP/FCSm	*h*-index	Quality
Univ A, 01	92	554	6.02	5.76	4.33	1.05	1.39	6	5
Univ A, 02	69	536	7.77	5.12	2.98	1.52	2.61	8	4
Univ A, 03	129	3780	29.3	17.2	11.86	1.7	2.47	17	5
Univ A, 04	80	725	9.06	8.06	6.25	1.12	1.45	7	4
Univ A, 05	188	1488	7.91	8.76	5.31	0.9	1.49	11	5
Univ A, 06	52	424	8.15	6.27	3.56	1.3	2.29	9	4
Univ A, 07	52	362	6.96	4.51	5.01	1.54	1.39	8	3
Univ A, 08	171	1646	9.63	6.45	4.36	1.49	2.21	13	5
Univ A, 09	132	2581	19.55	15.22	11.71	1.28	1.67	17	4
Univ A, 10	119	2815	23.66	22.23	14.25	1.06	1.66	17	4
Univ A, 11	141	1630	11.56	17.83	12.3	0.65	0.94	11	4
Univ A, 12	102	1025	10.05	10.48	7.18	0.96	1.4	10	5

Table 2 Pearson correlations (*lower triangle*) and Spearman rank correlations (*upper triangle*) among three citation indicators one peer-review based quality indicator

	CPP/JCSm	CPP/FCSm	*h*-index	Quality
CPP/JCSm		0.627*	0.057	−0.230
CPP/FCSm	0.783**		0.352	0.109
h-index	0.170	0.219		0.169
Quality	−0.133	0.156	0.151	

** $p < 0.01$; * $p < 0.05$

Van Raan (2006, p. 495) provided the Table 1[2]

Table 1 shows the results for 12 research groups in one university who published during this period 1,327 times, obtaining a total of 17,566 citations. The bibliometric indicators, the *h*-index, and the peer ratings are provided. In the latter, "5" indicates "excellent," "4" means "good," and "3" is classified as "satisfactory." Below "3" is not considered "satisfactory," but such a low rating did not occur in this set of data.

Table 2 shows the Pearson correlations (r) in the lower triangle and the Spearman rank correlations (ρ) in the upper triangle. As noted (cf. Van Raan 2006, p. 499), the *h*-index is also dependent on the number of publications while the CWTS-indicators are not. As could be expected, the two CWTS-indicators are highly correlated between themselves ($r = 0.783$; $p < 0.01$). However, the quality parameter Q is uncorrelated with any of these scientometric indicators. Thus, we may conclude that the indicators are *not* validated by this study despite the author's claim to the contrary.

Figure 1 shows the discriminating power of the *h*-index and the two indicators of CWTS (*CPP/JCSm* and *CPP/FCSm*) using the set provided in Table 1. We added error bars in order to show that the differences are contained within the margins of the standard errors of the measurement. Thus, none of the citation-based indicators is able to

[2] In footnotes 4 and 5 on p. 464, Van Raan (2006) explains the rationale for using different citation windows for the *h*-index and the CWTS indicators.

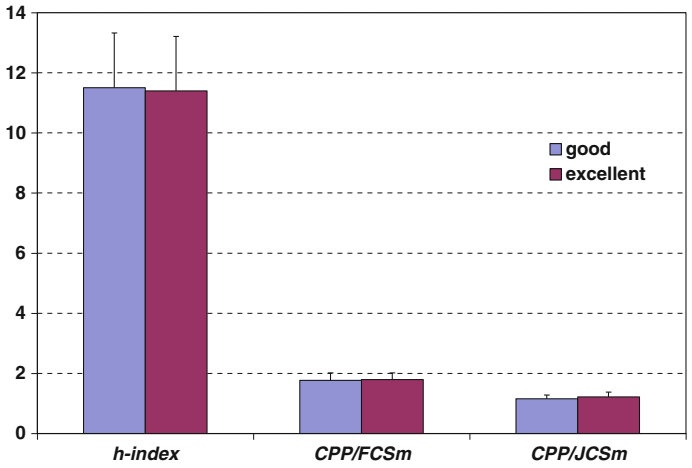

Fig. 1 Discrimination between "good" and "excellent" research using the *h*-index and the Leiden indicators CPP/JCSm and CPP/FCSm in the case of Table 1

discriminate between the categories "good" and "excellent" which were distinguished during the peer review.

In his Table 2, Van Raan (2006, p. 500) provided also aggregated data for the set of 147 research groups. In this table, the association between Q and h is significant (using χ^2, and $p < 0.05$), but not the association between Q and *CPP/FCSm* when testing $Q = 4$ against $Q = 5$ ($\chi^2 = 4.211$[3]; df = 2; $p = 0.112$). Thus, even at this aggregated level ($N = 147$), these results confirm the previous conclusion of Bornmann et al. (2010; cf. Van den Besselaar and Leydesdorff 2009) that the peer review systems and citation analysis are able to distinguish the tails of the distributions (low quality) from the high-end of the set, but perform poorly in distinguishing between excellent and good research to the extent that the correlation between evaluations based on these scientometric indicators or peer review can be negative (cf. Neufeld and von Ins 2011).

In Tables 19.3 and 19.4 at p. 243, Moed (2005) used this same data, but having access to the source data he added the larger set of similar results for biology and physics (whereas Table 1 above only contained the data for 12 research groups in chemistry in a single university among 147 chemistry groups at ten universities). By aggregating CPP/FCSm values also along the scale of "Citation impact classes," he can conclude (at p. 242) that "a very high citation impact discriminated very well between departments rated excellent or good and those receiving lower peer ratings, but it did not discriminate so well between good and excellent departments in the perception of peers."

This wording ("not so well") suggests a poor correlation, whereas we showed above that there was no correlation at the level of the smaller set of chemistry and using the values of CPP/FCSm before binning them into "Citation impact classes:" citation analysis is not always helpful in distinguishing between good and excellent research. Aggregation may inadvertently obscure the absence of correlations. Unfortunately, the selection between "excellent" and "good" is one of the policy contexts in which citation analysis is

[3] This value is Yates-corrected because of one value smaller than five. Without this correction: $\chi^2 = 5.559$; df = 2; $p = 0.062$.

used; for example, in rankings and funding schemes (e.g., Bornmann et al. 2010; Halffman and Leydesdorff 2010; Geuna and Martin 2003).

In summary, we argue that the industrial character of citation analysis for evaluative purposes has hidden technical flaws in these measurements because of a lack of openness about the data and therefore critical discussion in academia. Notwithstanding their prevailing use in research evaluation and strategic decision-making, the statistical analysis of this scientometric data, for example, supports the claim of the criticizers (e.g., MacRoberts and MacRoberts 2010) that citation analysis hitherto cannot legitimate the strategic selection of excellence.

References

Bornmann, L., Leydesdorff, L., & Van den Besselaar, P. (2010). A meta-evaluation of scientific research proposals: Different ways of comparing rejected to awarded applications. *Journal of Informetrics, 4*(3), 211–220.

Geuna, A., & Martin, B. R. (2003). University research evaluation and funding: An international comparison. *Minerva, 41*(4), 277–304.

Halffman, W., & Leydesdorff, L. (2010). Is inequality among universities increasing? Gini coefficients and the elusive rise of elite universities. *Minerva, 48*(1), 55–72.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the USA, 102*(46), 16569–16572.

Leydesdorff, L. (2008). *Caveats* for the use of citation indicators in research and journal evaluation. *Journal of the American Society for Information Science and Technology, 59*(2), 278–287.

Leydesdorff, L., & Opthof, T. (2010). Normalization at the field level: Fractional counting of citations. *Journal of Informetrics, 4*(4), 644–646.

Leydesdorff, L., & Opthof, T. (2011). Remaining problems with the "New Crown Indicator" (MNCS) of the CWTS. *Journal of Informetrics, 5*(1), 224–225.

MacRoberts, M. H., & MacRoberts, B. R. (1987). Another test of the normative theory of citing. *Journal of the American Society for Information Science, 16*, 151–172.

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

MacRoberts, M. H., & MacRoberts, B. R. (2010). Problems of citation analysis: A study of uncited and seldom-cited influences. *Journal of the American Society for Information Science and Technology, 61*(1), 1–12.

Moed, H. F. (2005). *Citation analysis in research evaluation*. Dordrecht: Springer.

Moed, H. F., De Bruin, R. E., & Van Leeuwen, T. N. (1995). New bibliometric tools for the assessment of national research performance: Database description, overview of indicators and first applications. *Scientometrics, 33*(3), 381–422.

Narin, F. (1976). *Evaluative bibliometrics: The use of publication and citation analysis in the evaluation of scientific activity*. Washington, DC: National Science Foundation.

Neufeld, J., & von Ins, M. (2011). Informed peer review and uninformed bibliometrics? *Research Evaluation, 20*(1), 31–46.

Opthof, T., & Leydesdorff, L. (2010). *Caveats* for the journal and field normalizations in the CWTS ("Leiden") evaluations of research performance. *Journal of Informetrics, 4*(3), 423–430.

Schubert, A., & Braun, T. (1986). Relative indicators and relational charts for comparative assessment of publication output and citation impact. *Scientometrics, 9*(5), 281–291.

Van den Besselaar, P., & Leydesdorff, L. (2009). Past performance, peer review, and project selection: A case study in the social and behavioral sciences. *Research Evaluation, 18*(4), 273–288.

Van Raan, A. F. J. (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. (1986). Evaluation of some methods for the relative assessment of scientific publications. *Scientometrics, 10*(3), 157–177.

Vinkler, P. (1996). Model for quantitative selection of relative scientometric impact indicators. *Scientometrics, 36*(2), 223–236.

VSNU. (2002). *Chemistry and chemical engineering*. VSNU Series 'Assessment of Research Quality.' Utrecht: VSNU. ISBN 90 5588 4979.

Waltman, L., van Eck, N. J., van Leeuwen, T. N., Visser, M. S., & Van Raan, A. F. J. (2011a). Towards a new crown indicator: An empirical analysis. *Scientometrics, 87*, 467–481.

Waltman, L., Van Eck, N. J., Van Leeuwen, T. N., Visser, M. S., & Van Raan, A. F. J. (2011b). Towards a new crown indicator: Some theoretical considerations. *Journal of Informetrics, 5*(1), 37–47.