



Contents lists available at ScienceDirect

Journal of Informetrics

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

Letter to the Editor

Towards an ideal method of measuring research performance: Some comments to the Opthof and Leydesdorff (2010) paper

The paper of Opthof and Leydesdorff (2010) proposes some changes in the way, in which the Centre for Science and Technology Studies (CWTS, Leiden, The Netherlands) uses citation counts for the evaluation and comparison of research performance. These changes focus on three points concerning the journal- and field-normalization of citations: (1) The use of field classifications developed by discipline-oriented databases (e.g., Chemical Abstracts, CA, provided by Chemical Abstracts Services, CAS, Columbus, OH, USA) compared to Thomson Reuters (Philadelphia, PA, USA) journal-based subject categories. (2) Each paper in a paper set becomes field-normalized, before an average value over the field-normalized values is calculated. CWTS calculates the average value over all papers in the set and then field-normalizes the average value. (3) The avoidance of the arithmetic average. In the face of non-normal distributed citation data, the arithmetic mean value is not appropriate as a measure of central tendency. Since many years, the method of CWTS has been defined as being the standard in the evaluative bibliometry. In general, the proposals of Opthof and Leydesdorff (2010) are highly interesting and very important although they are in part criticizable. Meanwhile, the paper of Opthof and Leydesdorff (2010) has been discussed in the Sigmetrics mailing list and an answer of the CWTS was written (Van Raan et al., 2010).

In this letter to the editor, the proposals of Opthof and Leydesdorff (2010) are took up and extended. In their paper, the authors had already referred to two of our papers (Bornmann & Daniel, 2009; Bornmann, Mutz, Neuhaus, & Daniel, 2008). Below, some points are noted to advance towards an ideal method to compare the citation performance of different research groups:

1. In general, only full length articles, letters, notes, communications and reviews should be considered in an evaluative citation analysis.
2. To standardize citation counts the total number of citations should be gathered for a fixed time window of several years after the publication year (at least 3 years). "Fixed citation windows are a standard method in bibliometric analysis, in order to give equal time spans for citation to articles published in different years, or at different times in the same year" (Craig et al., 2007, p. 243). Both, the citation counts for the papers of the research groups as well as the citation counts for those papers the reference standards are based on should be time fixed.
3. In order to compare the citation counts of papers for different research groups, no arithmetic mean should be calculated. Opthof and Leydesdorff (2010) propose to avoid arithmetic means, but they do not waive it completely. They still use mean citation counts as reference standards. In the face of non-normal distributed citation data, the arithmetic mean value can give a distorted picture of the kind of distribution (Bornmann et al., 2008), "and it is a rather crude statistic" (Joint Committee on Quantitative Assessment of Research, 2008, p. 2). As the distribution of citation counts is usually right-skewed, distributed according to a power law (Joint Committee on Quantitative Assessment of Research, 2008), arithmetic average citation rates mainly show where papers with high citation counts are to be found. According to Evidence Ltd., 2007 "where bibliometric data must stand-alone, they should be treated as distributions and not as averages" (p. 10). Percentiles, instead of the arithmetic average, should be used in an evaluative citation analysis. In educational and psychological testing, percentile rank scores are widely used as a standard for comparison, in order to judge a person's test scores (e.g., intelligence test scores) based on a comparison with the percentiles of a calibrated sample (Bornmann, Mutz, & Daniel, 2007). Particularly in bibliometric analysis the use of percentiles is very advantageous (see Evidence Ltd., 2007; Plomp, 1990), because no assumptions have to be made as to the distribution of citations.
4. The CWTS uses reference standards based on journal classification schemes (Neuhaus & Daniel, 2009). Each journal is classified as a whole to one or several subject categories. The limitations of the CWTS method become obvious in the case of multidisciplinary journals such as *Nature* or *Science* and highly specialized fields of research (Opthof & Leydesdorff, 2010). The difficulty on the one hand are papers that appear in multidisciplinary journals because they cannot be assigned exclusively to one field and on the other hand highly specialized fields to which no adequate reference values exist. To overcome the limitations of journal classification schemes, Neuhaus and Daniel (2009) propose an alternative reference

Table 1

Absolute (abs) and relative (%) number of papers published by three research groups which were categorized into six percentile rank classes (fictitious data).

Percentile rank class	Research group 1		Research group 2		Research group 3	
	abs	%	abs	%	abs	%
<50th	43	28	17	10 ⁻	68	27
50th	22	14	28	16	43	17
75th	33	21 ⁺	19	11	27	11
90th	21	13	39	23	36	15
95th	23	15	22	13	40	16
99th	14	9 ⁻	45	27 ⁺	36	14
Total	156	100	170	100	250	100

Notes. $\chi^2_{10} = 48.1$, $P < .05$, Cramér's $V = .20$. A plus or minus sign indicates a lack of fit for independence in that cell (Agresti, 2002).

standard that is based on a paper-by-paper basis (see also Neuhaus, Marx, & Daniel, 2009). In contrast to a reference standard based on journal sets (where all papers in a journal are assigned to one and the same field) for the alternative reference standard every paper is associated with a single principal (sub-)field entry which accentuates the most important aspect of the work (see here also Kurtz & Henneken, 2007; Pendlebury, 2008). "In discipline-oriented databases such as *Chemical Abstracts*, *MEDLINE*, or *INSPEC*, fields and subfields can be identified by means of a structured subject classification scheme" (Neuhaus & Daniel, 2009, p. 222). The Medline database of the National Library of Medicine (NLM, Bethesda, MD, USA) provides "detailed and high-quality medical subject headings [MeSH terms] . . . for field delimitations" (Strotmann & Zhao, 2010, p. 196). For CA, CAS categorizes chemical publications into 80 different subject areas (chemical fields, called "sections"). Every publication is assigned to a single principal entry that clearly demonstrates the most important aspect of the chemical work (Daniel, 1993). For an evaluative bibliometric study, it is not necessary that the discipline-oriented database provides cited-reference information in addition to the field classification scheme. The study of Strotmann and Zhao (2010) shows how information from multidisciplinary citation indexes (e.g., cited-reference information from Scopus) can be matched with information from discipline-oriented databases (e.g., MeSH terms from Medline). Following the proposal of Opthof and Leydesdorff (2010) each paper in a paper set of a research group should be field-normalized with a matching reference standard. To generate the reference standard for a paper in question all papers published in the same year, with the same document type and belonging to the same field (defined by a discipline-oriented database) are categorized into six percentile rank classes: 99th, 95th, 90th, 75th, 50th, and <50th (following the approach of the National Science Board, 2010). Through the use of the citation limits that define these classes the paper in question can be assigned to one of the six citation impact classes. This procedure is repeated for all papers published by a research group.

The results of an evaluative citation analysis as proposed here would look like the findings displayed in Table 1. The table reports the results calculated with fictitious bibliometric data for three research groups. The papers of the groups were categorized into six percentile rank classes. Table 1 shows the distribution of the papers over the classes. With 9% (research group 1), 27% (research group 2) and 14% (research group 3) all three groups have published significantly more top-level papers than the expected value of 1% (99th percentile). Furthermore, with 28% (research group 1), 10% (research group 2) and 27% (research group 3) the three groups have less papers in the lowest percentile rank class than the expected value of >50% (<50th percentile). Compared to calibrated samples, the three groups show a better citation performance than one would normally expect for the fields where they publish. Citation impact differences between the three groups are only meaningful if they are statistically significant. As the results of the chi-square test printed in the notes of Table 1 show, the differences are in fact statistically significant and therefore meaningful. Indications for performance differences between the three groups at certain impact classes (e.g., the 99th percentile) are provided by the calculation of standardized Pearson residuals. Residuals that exceed about 2 or 3 (in absolute value) suggest the cell having a lack of fit for independence (Agresti, 2002). In Table 1, a plus or minus sign indicates this lack of fit: Compared to the other groups, the citation performance of group 2 is characterized by a high percentage of papers belonging to the 99th percentile (27%). In contrast, the performance of research group 1 is characterized by a low percentage of papers in this class (9%).

Since the result of the statistical significance test is dependent on sample size and "statistical significance does not mean real life importance" (Conroy, 2002, p. 290), it is the strength of the association between research group and citation performance that is additionally interesting and important for interpreting the empirical finding in the table. In order to calculate strength, we have to employ a measure of association, i.e., Cramér's V coefficient (Cramér, 1980). According to Kline (2004), Cramér's V "is probably the best known measure of association for contingency tables" (p. 151). A Cramér's V value of .20 indicated in the note of Table 1 points out a medium effect size for the association between citation performance and research group. A medium effect size can be interpreted as typical in behavioral research (see here Kraemer et al., 2003).

This example for doing an evaluative citation analysis with fictitious data should demonstrate the advantages of the procedure proposed here for measuring and comparing research performance which is partly based on the proposals of Opthof and Leydesdorff (2010). Although the proposed method may entail many advantages over the present standards in evaluative bibliometrics, it is an important task for prospective research to further work towards an ideal way of measuring

research performance: “Measurement of research excellence and quality is an issue that has increasingly interested governments, universities, and funding bodies as measures of accountability and quality are sought” (Steele, Butler, & Kingsley, 2006, p. 278).

References

- Agresti, A. (2002). *Categorical data analysis*. Hoboken, NJ, USA: John Wiley & Sons, Inc.
- Bornmann, L., & Daniel, H.-D. (2009). Universality of citation distributions. A validation of Radicchi et al.'s relative indicator $c_f = c/c_0$ at the micro level using data from chemistry. *Journal of the American Society for Information Science and Technology*, 60(8), 1664–1670.
- Bornmann, L., Mutz, R., & Daniel, H.-D. (2007). The *b* index as a measure of scientific excellence. A promising supplement to the *h* index. *Cybermetrics*, 11(1) (paper 6).
- Bornmann, L., Mutz, R., Neuhaus, C., & Daniel, H.-D. (2008). Use of citation counts for research evaluation: standards of good practice for analyzing bibliometric data and presenting and interpreting results. *Ethics in Science and Environmental Politics*, 8, 93–102. doi:10.3354/esep00084
- Conroy, R. M. (2002). Choosing an appropriate real-life measure of effect size: the case of a continuous predictor and a binary outcome. *The Stata Journal*, 2(3), 290–295.
- Craig, I. D., Plume, A. M., McVeigh, M. E., Pringle, J., & Amin, M. (2007). Do open access articles have greater citation impact? A critical review of the literature. *Journal of Informetrics*, 1(3), 239–248. doi:10.1016/j.joi.2007.04.001
- Cramér, H. (1980). *Mathematical methods of statistics*. Princeton, NJ, USA: Princeton University Press.
- Daniel, H.-D. (1993). *Guardians of science. Fairness and reliability of peer review*. Weinheim, Germany: Wiley-VCH.
- Evidence Ltd. (2007). *The use of bibliometrics to measure research quality in UK higher education institutions*. London, UK: Universities UK.
- Joint Committee on Quantitative Assessment of and Research. (2008). *Citation statistics. A report from the International Mathematical Union (IMU) in cooperation with the International Council of Industrial and Applied Mathematics (CIAM) and the Institute of Mathematical Statistics (IMS)*. Berlin, Germany: International Mathematical Union (IMU).
- Kline, R. B. (2004). *Beyond significance testing: reforming data analysis methods in behavioral research*. Washington, DC, USA: American Psychological Association.
- Kraemer, H. C., Morgan, G. A., Leech, N. L., Gliner, J. A., Vaske, J. J., & Harmon, R. J. (2003). Measures of clinical significance. *Journal of the American Academy of Child and Adolescent Psychiatry*, 42(12), 1524–1529. doi:10.1097/01.chi.0000091507.46853.d1
- Kurtz, M. J., & Henneken, E. A. (2007). Open access does not increase citations for research articles. *The Astrophysical Journal*, Retrieved September 10, 2007, from <http://arxiv.org/abs/0709.0896>
- National Science Board. (2010). *Science and engineering indicators 2010, appendix tables*. Arlington, VA, USA: National Science Foundation.
- Neuhaus, C., & Daniel, H.-D. (2009). A new reference standard for citation analysis in chemistry and related fields based on the sections of Chemical Abstracts. *Scientometrics*, 78(2), 219–229.
- Neuhaus, C., Marx, W., & Daniel, H.-D. (2009). The publication and citation impact profiles of *Angewandte Chemie* and the *Journal of the American Chemical Society* based on the sections of *Chemical Abstracts*: a case study on the limitations of the Journal Impact Factor. *Journal of the American Society for Information Science and Technology*, 60(1), 176–183. doi:10.1002/asi.20960
- Ophof, 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.
- Pendlebury, D. A. (2008). *Using bibliometrics in evaluating research*. Philadelphia, PA, USA: Research Department, Thomson Scientific.
- Plomp, R. (1990). The significance of the number of highly cited papers as an indicator of scientific prolificacy. *Scientometrics*, 19(3–4), 185–197.
- Steele, C., Butler, L., & Kingsley, D. (2006). The Publishing imperative: the pervasive influence of publication metrics. *Learned Publishing*, 19(4), 277–290.
- Strotmann, A., & Zhao, D. (2010). Combining commercial citation indexes and open-access bibliographic databases to delimit highly interdisciplinary research fields for citation analysis. *Journal of Informetrics*, 4(2), 194–200. doi:10.1016/j.joi.2009.12.001
- Van Raan, A. F. J., Van Leeuwen, T. N., Visser, M. S., Van Eck, N. J., & Waltman, L. (2010). Rivals for the crown: reply to Ophof and Leydesdorff. *Journal of Informetrics*, 4(3), 431–435.

Lutz Bornmann*

ETH Zurich, Professorship for Social Psychology and Research on Higher Education,
Zähringerstr. 24, CH-8092 Zurich, Switzerland

*Tel.: +41 44 632 48 25; fax: +41 44 632 12 83.
E-mail address: bornmann@gess.ethz.ch

12 April 2010