Journal of Informetrics 8 (2014) 273–275

Contents lists available at ScienceDirect

Journal of Informetrics

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

Letter to the Editor

On the meaningful and non-meaningful use of reference sets in bibliometrics

In a paper published recently, Kaur, Radicchi, and Menczer (2013) used data from the Scholarometer (http://scholarometer.indiana.edu/) to examine the effectiveness of various metrics, such as the h index (Hirsch, 2005) and the new crown indicator (Lundberg, 2007; Opthof & Leydesdorff, 2010; Waltman, van Eck, van Leeuwen, Visser, & van Raan, 2011a; Waltman, van Eck, van Leeuwen, Visser, & van Raan, 2011b) in generating field-normalized citation scores. While the subject of field-normalization in bibliometrics has up to now been discussed primarily at the paper level (Bornmann & Leydesdorff, 2013), Kaur et al. (2013) have been looking at the author level: which metric allows an effective method of field-normalization with which scientists can be compared across different fields? There have already been numerous calls for benchmarks for comparative assessment in order to evaluate individual scientists (Garfield, 1979; Kreiman & Maunsell, 2011). Something that is very difficult to implement for individuals (Coleman, Bolumole, & Frankel, 2012; El Emam, Arbuckle, Jonker, & Anderson, 2012) can be achieved at the institutional level with the data from the rankings, as Bornmann and de Moya Anegón (2014) have shown in their study (provided that one accepts the reasoned decisions which are made for constructing the rankings).

We would like to use the paper by Kaur et al. (2013) as a starting point for an investigation into an appropriate method of field-normalization (at individual scientist level). In bibliometrics, in order to allow cross-disciplinary comparisons of citation impact at the level of individual papers, a reference set made up of all the papers from the same field (and the same publication year) is compiled for each paper. One can expect that the Web of Science (WoS, Thomson Reuters) and Scopus (Elsevier) have a good coverage of the literature in the natural and life sciences (Mahdi, d'Este, & Neely, 2008). Only by taking all the comparable papers into account is it possible to ensure that the measurement of the impact of the paper in question is valid compared to similar papers (Bornmann & Marx, 2013a). If only some of the total numbers of papers are used in the reference sets, there is no comparison with the relevant reference sets.

In their study, Kaur et al. (2013) use bibliometric data for scientists who have used the Scholarometer tool to normalize metrics at the level of individual scientists. As we can assume that not all scientists worldwide use this tool (nor a sample which can be interpreted as meaningful, such as all scientists with at least one paper in WoS), it is not possible to generate meaningful and valid reference data at the level of scientists on this basis. In order to test the various metrics in their study, the authors would have had to normalize each scientist recorded in Scholarometer with an appropriate valid reference set (including all scientists). Only then would they have been able to verify the effectiveness of the metrics using the example of scientists from Scholarometer.

Where the new crown indicator is concerned, which Kaur et al. (2013) included in their comparative study along with others, there is an additional problem in that it was obviously also calculated with Scholarometer data. The expected citation rate for a publication is therefore not – as is a currently standard in bibliometrics – calculated over the impact of all the publications in a subject category of the Web of Science (WoS, Thomson Reuters) or in a Scopus (Elsevier) subject area and a publication year, but over the impact of a selection of publications which users of the Scholarometer have entered quite randomly and assigned to certain disciplines. Can we call this a valid reference set? We would say not.

Bibliometric data used to evaluate research on the level of individual scientists is highly critical data and should therefore be compiled very carefully (Bornmann & Marx, 2013b; Marx & Bornmann, 2014). In the Scholarometer, users can enter names of scientists as they wish, assign these scientists to certain disciplines, and compile their publication sets. Even though there are processes implemented in the Scholarometer which are supposed to prevent serious misclassifications (Kaur et al., 2012), one can assume that these assignments are not high-quality. However, we need high-quality data when we put reference sets together and use them for the evaluation of research. Indeed, the database operator Chemical Abstracts Services, for example, employs a number of highly specialised people to assign individual publications in chemistry and its related fields to specific subject categories (Bornmann, Marx, & Barth, 2013; Bornmann, Mutz, Marx, Schier, & Daniel, 2011). Thomson Reuters (WoS) and Elsevier (Scopus) have addressed the problem of subject matter classification by assigning journals (and not individual papers) to subject categories. Despite much criticism (Bornmann, Mutz, Neuhaus, & Daniel, 2008b; Rafols & Leydesdorff, 2009), this classification is currently used as a standard in bibliometrics.

1751-1577/$ – see front matter
http://dx.doi.org/10.1016/j.joi.2013.12.006

Unlike other studies which have looked at the normalization of citation impact, Kaur et al. (2013) normalize on the level of individual scientists. If one normalizes at the level of individual authors, it is essential to take into account not only the publication year and the field of the papers, but also the academic age of the scientist (Bornmann & Marx, 2013b). Biochemists in the Scholarometer database may perform better than biologists on average, only because on average the biochemists who used the tool were older than the biologists. The differences in performance which Kaur et al. (2013) show in Fig. 1, for example, are not necessarily related to the subject. So when working at the level of individual authors, one should also normalize for the year in each case. It is possible to do this by dividing the metric by the academic age. With the *m* quotient, which was proposed by Hirsch (2005), *h* is divided by the number of years since the first publication. There is another option: to include only scientists of a specific academic age in a reference set. For example, one could include all the scientists whose first publication appeared within a certain period.

The final points that we would like to address concerns (1) the metrics which Kaur et al. (2013) included in their tool and (2) the data base used (Google Scholar, GS). (1) Strictly speaking, the metrics are not directly comparable with each other. As a number of studies of the *h* index and its variants have shown, primarily these studies measure output and not impact (Bornmann, Mutz, & Daniel, 2008; Bornmann, Mutz, Daniel, 2009; Bornmann, Mutz, Daniel, Wallon, et al., 2009). It is therefore not possible to compare them to pure citation-based measures such as the new crown indicator. (2) We do not advise to use GS as a sole basis for a bibliometric analysis. Several studies have pointed out that GS has numerous deficiencies for research evaluation (Bornmann, Marx, et al., 2009; García-Pérez, 2010; Jacso, 2009, 2010). For Jacso (2008) GS "does a really horrible job matching cited and citing references" and "often can't tell apart a page number from a publication year, part of the title of a book from a journal name, and dumps at you absurd data." Meho and Yang (2007) conclude that overall, GS is "not conducive for large-scale comparative citation analyses" (p. 579).

References

Bornmann, L., & de Moya Anegón, F. (2014). What proportion of excellent papers makes an institution to one of the best worldwide? Specifying thresholds for the interpretation of the results of the SCImago Institutions Ranking and the Leiden Ranking. *Journal of the American Society for Information Science and Technology* (in press)

Bornmann, L., & Leydesdorff, L. (2013). The validation of (advanced) bibliometric indicators through peer assessments: A comparative study using data from InCites and F1000. *Journal of Informetrics*, *7*(2), 286–291. http://dx.doi.org/10.1016/j.joi.2012.12.003

Bornmann, L., & Marx, W. (2013a). How good is research really? Measuring the citation impact of publications with percentiles increases correct assessments and fair comparisons. *EMBO Reports*, *14*(3), 226–230. http://dx.doi.org/10.1038/embor.2013.9

Bornmann, L., & Marx, W. (2013b). How to evaluate individual researchers working in the natural and life sciences meaningfully? A proposal of methods based on percentiles of citations. *Scientometrics*, 1–23. http://dx.doi.org/10.1007/s11192-013-1161-y

Bornmann, L., Marx, W., & Barth, A. (2013). The normalization of citation counts based on classification systems. *Publications*, *1*(2), 78–86.

Bornmann, L., Marx, W., Schier, H., Rahm, E., Thor, A., & Daniel, H. D. (2009). Convergent validity of bibliometric Google Scholar data in the field of chemistry. Citation counts for papers that were accepted by Angewandte Chemie International Edition or rejected but published elsewhere, using Google Scholar, Science Citation Index, Scopus, and Chemical Abstract. *Journal of Informetrics*, *3*(1), 27–35. http://dx.doi.org/10.1016/j.joi.2008.11.001

Bornmann, L., Mutz, R., & 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 American Society for Information Science and Technology*, *59*(5), 830–837. http://dx.doi.org/10.1002/asi.20806

Bornmann, L., Mutz, R., & Daniel, H.-D. (2009). Do we need the *h* index and its variants in addition to standard bibliometric measures? *Journal of the American Society For Information Science and Technology*, *60*(6), 1286–1289. http://dx.doi.org/10.1002/asi.21016

Bornmann, L., Mutz, R., Daniel, H.-D., Wallon, G., & Ledin, A. (2009). Are there really two types of *h* index variants? A validation study by using molecular life sciences data. *Research Evaluation*, *18*(3), 185–190. http://dx.doi.org/10.3152/095820209X466883

Bornmann, L., Mutz, R., Marx, W., Schier, H., & Daniel, H.-D. (2011). A multilevel modelling approach to investigating the predictive validity of editorial decisions: do the editors of a high profile journal select manuscripts that are highly cited after publication? *Journal of the Royal Statistical Society Series a-Statistics in Society*, *174*, 857–879. http://dx.doi.org/10.1111/j.1467-985X.2011.00689.x

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. http://dx.doi.org/10.3354/esep00084

Coleman, B. J., Bolumole, Y. A., & Frankel, R. (2012). Benchmarking individual publication productivity in logistics. *Transportation Journal*, *51*(2), 164–196.

El Emam, K., Arbuckle, L., Jonker, E., & Anderson, K. (2012). Two *h*-index benchmarks for evaluating the publication performance of medical informatics researchers. *Journal of Medical Internet Research*, *14*(5) http://dx.doi.org/10.2196/jmir.2177

García-Pérez, M. A. (2010). Accuracy and completeness of publication and citation records in the Web of Science, PsycINFO, and Google Scholar: A case study for the computation of h indices in Psychology. *Journal of the American Society for Information Science and Technology*, *61*(10), 2070–2085. http://dx.doi.org/10.1002/asi.21372

Garfield, E. (1979). *Citation indexing – Its theory and application in science, technology, and humanities*. New York, NY, USA: John Wiley & Sons, Ltd.

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*(46), 16569–16572. http://dx.doi.org/10.1073/pnas.0507655102

Jacso, P. (2008). *Google Scholar and The Scientist*. Retrieved from http://www2.hawaii.edu/~jacso/extra/gs/

Jacso, P. (2009). Google Scholar's ghost authors. *Library Journal*, *134*(18), 26–27.

Jacso, P. (2010). Metadata mega mess in Google Scholar. *Online Information Review*, *34*(1), 175–191. http://dx.doi.org/10.1108/14684521011024191

Kaur, J., Hoang, D. T., Sun, X., Possamai, L., JafariAsbagh, M., Patil, S., et al. (2012). Scholarometer: a social framework for analyzing impact across disciplines. *PLoS ONE*, *7*(9), e43235. http://dx.doi.org/10.1371/journal.pone.0043235

Kaur, J., Radicchi, F., & Menczer, F. (2013). Universality of scholarly impact metrics. *Journal of Informetrics*, *7*(4), 924–932. http://dx.doi.org/10.1016/j.joi.2013.09.002

Kreiman, G., & Maunsell, J. H. R. (2011). Nine criteria for a measure of scientific output. *Frontiers in Computational Neuroscience*, *5* http://dx.doi.org/10.3389/fncom.2011.00048

Lundberg, J. (2007). Lifting the crown – Citation *z*-score. *Journal of Informetrics*, *1*(2), 145–154.

Mahdi, S., d'Este, P., & Neely, A. D. (2008). *Citation counts: are they good predictors of RAE scores? A bibliometric analysis of RAE 2001*. London, UK: Advanced Institute of Management Research.

Marx, W., & Bornmann, L. (2014). On the problems of dealing with bibliometric data. *Journal of the American Society for Information Sciences and Technology* (in press).

Meho, L. I., & Yang, K. (2007). Fusion approach to citation-based quality assessment. In D. Torres-Salinas, & H. F. Moed (Eds.), *Proceedings of the 11th Conference of the International Society for Scientometrics and Informetrics* (2) (pp. 568–581). Madrid, Spain: Spanish Research Council (CSIC).

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.

Rafols, I., & Leydesdorff, L. (2009). Content-based and algorithmic classifications of journals: Perspectives on the dynamics of scientific communication and indexer effects. *Journal of the American Society for Information Science and Technology*, *60*(9), 1823–1835.

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*(3), 467–481. http://dx.doi.org/10.1007/s11192-011-0354-5

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. http://dx.doi.org/10.1016/j.joi.2010.08.001

Lutz Bornmann*
Division for Science and Innovation Studies, Administrative Headquarters of the Max Planck Society, Hofgartenstr. 8, 80539 Munich, Germany

Loet Leydesdorff
Amsterdam School of Communication Research (ASCoR), University of Amsterdam, Kloveniersburgwal 48, 1012 CX Amsterdam, The Netherlands

* Corresponding author. Tel. +49 89 2108 1265.
E-mail addresses: bornmann@gv.mpg.de (L. Bornmann), loet@leydesdorff.net (L. Leydesdorff)

5 November 2013

Available online 11 January 2014