



Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: [www.elsevier.com/locate/joi](http://www.elsevier.com/locate/joi)

## Letter to the Editor

**Further steps towards an ideal method of measuring citation performance: The avoidance of citation (ratio) averages in field-normalization**

Since many years the so called crown indicator (CPP/FCS<sub>m</sub>) for field normalization of the Centre for Science and Technology Studies (CWTS, Leiden, The Netherlands) has been defined as being the standard in the evaluative bibliometric practice in many different contexts. The publication of the paper by Opthof and Leydesdorff (2010) was a starting point in the field of evaluative bibliometrics to challenge this CWTS standard indicator for evaluative purposes. Meanwhile, the paper of Opthof and Leydesdorff (2010) has been extensively discussed (Van Raan, Van Leeuwen, Visser, Van Eck, & Waltman, 2010), e.g., during the Science and Technology Indicators conference 2010 in Leiden (Anon, 2010), and a new crown indicator was presented by CWTS: the mean normalized citation score (MNCS) (Waltman, van Eck, van Leeuwen, Visser, & van Raan, 2011). At its heart both, the previous and new indicator differentiates as follows: whereas for the previous crown indicator the average citation rate over all papers in a publication set is calculated and then the citation rate is field-normalized, for the new crown indicator each paper's citation impact in a paper set becomes field-normalized, before an average value (harmonic average) over the field-normalized impact values is calculated.

Not only for the previous but also for the new crown indicator the disadvantage of resting on the arithmetic average exists (Leydesdorff & Opthof, in press): the mean citation impact values calculated for different fields are arithmetic averages and the crown indicators rest on arithmetic averages of ratios or ratios of arithmetic averages. A number of publications has pointed out that in the face of non-normal distributed citation data, the arithmetic mean value is not appropriate as a measure of central tendency. It can give a distorted picture of the kind of distribution (Bornmann, Mutz, Neuhaus, & Daniel, 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). What is more, the calculation of ratios runs into serious problems regarding the interpretation of citation impact as low or high (or excellent). The interpretation is more or less arbitrary without using reference distributions. Since many years, the evaluation of the citation performance of research groups as far below (<0.5) or far above (>1.5) the international citation impact standard has been based on cut-off points developed by personal experiences (see here van Raan, 2005).

In the following we will present an extension of the proposals of Bornmann (2010) for an improved practise of field-normalized citation performance measurement. The extension is intended to calculate a single measure for the citation impact of a group of scientists that is not based on the arithmetic average but uses reference distributions. The measure allows – similar to the previous and new crown indicator – to compare groups of scientists by using one single number. The proposals of Bornmann (2010) are based on the calculation of percentiles. The use of percentiles for citation data is very advantageous, because no assumptions have to be made as to the underlying distribution of citations (Boyack, 2004). With percentiles each paper in a paper set of a research group can 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 impact classes: 99th – top 1%, 95th, 90th, 75th, 50th, and <50th – bottom 50% (following the approach of the National Science Board, 2010) (see here also Bornmann, de Moya-Anegón, & Leydesdorff, 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 use of the percentile for each paper instead of the corresponding percentile impact class might be more preferable (e.g., resulting in a higher power), but this is a very cumbersome and expensive task for a bigger publication set.).

Bornmann (2010) presents the results of an evaluative citation analysis calculated with fictitious bibliometric data for three research groups. Table 1 shows this data with some additional numbers. As the table reveals the papers of the groups were categorized into six percentile impact classes (99th – top 1%, 95th, 90th, 75th, 50th, and <50th – bottom 50%). First of all,

**Table 1**

Absolute (*n*), relative (*p*), and cumulative (*cump*) frequencies as well as expected values (*EV*) for papers published by three research groups which were categorized into six percentile impact classes (fictitious data).

Percentile impact class	Numbering of classes ( <i>X</i> )	Research group 1				Research Group 2				Research group 3			
		<i>n</i>	<i>p</i>	<i>cump</i>	<i>EV</i>	<i>n</i>	<i>p</i>	<i>cump</i>	<i>EV</i>	<i>n</i>	<i>p</i>	<i>cump</i>	<i>EV</i>
<50th (bottom 50%)	1	43	0.3	0.3	0.3	17	0.1-	0.1	0.1	68	0.3	0.3	0.3
50th	2	22	0.1	0.7	0.2	28	0.2	0.9	0.4	43	0.2	0.7	0.4
75th	3	33	0.2+	0.6	0.6	19	0.1	0.7	0.3	27	0.1	0.5	0.3
90th	4	21	0.1	0.4	0.4	39	0.2	0.6	0.8	36	0.1	0.4	0.4
95th	5	23	0.2	0.3	1.0	22	0.1	0.4	0.5	40	0.2	0.3	1.0
99th (top 1%)	6	14	0.1-	0.1	0.6	45	0.3+	0.3	1.8	36	0.1	0.1	0.6
Total		156	1.0		3.1	170	1.0		3.9	250	1.0		3.0

Notes: *cump*, relative frequency of papers to be in percentile class *X* or higher (the latter except percentile class 1).  $\chi^2(10, N=576)=48.1, p<.05$ , Cramér's *V* = .20. A plus or minus sign indicates a lack of fit for independence in that cell (Agresti, 2002).

the use of these classes as a reference distribution being expected for a research group with a medium performance allows the evaluation of each research group's citation impact in Table 1 for its own. In column *p* the frequencies (or probabilities) of papers falling in a certain percentile impact class (and not in the next higher class) are listed. The column *cump* indicates the frequency of papers of a research group which falls into a certain percentile impact class or in all higher classes (the latter except the lowest class). For a research group with a medium performance it can be assumed that the distribution of papers across the percentile impact classes corresponds to the proportions given by the classes: 50% of the papers of this medium group are in the lowest (bottom 50%), 10% in the 90th, and 1% in the highest percentile impact class (top 1%). In our fictitious example (see Table 1) only 30% (*p*=0.3) of all 156 papers published by research group 1 fall into the bottom 50% impact class, but 70% in the classes higher than 50%. 10% (*p*=0.1) of all papers fall into the 99th class (top 1%) and thus, 10 times more than in the reference set. As this comparison with the reference proportions given by the percentile impact classes shows the citation impact of research group 1 is much higher than being expected for a medium performance group. This is also true for research groups 2 (*p*=0.3) and 3 (*p*=0.1).

With the statistics presented below the performance differences between the three groups can be tested on significance. The results of the  $\chi^2$  test printed in the legend of Table 1 show that the citation impact differences between the groups are 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. 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 number of papers belonging to the top 1% (*p*=0.3). In contrast, the performance of research group 1 is characterized by a low number of papers in this class (*p*=0.1), albeit still substantially higher than the expected frequency of 1%. A Cramér's *V* value of .20 (see the legend 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 social science studies (see here Kraemer et al., 2003). These results are presented in Bornmann (2010). The statistics used for the fictitious data indicate that the research groups differ meaningful in citation impact – within certain citation impact classes and over all classes.

In addition to these statistics the calculation of an expected value (*EV*) is proposed (see Table 1) which allows a comparison of the citation performance of different research groups by using one single measure. It is the advantage of this value against measures like the former and new crown indicator not only to abandon the arithmetic average, but also to give thresholds for the minimal, medium and maximal possible citation performance. The *EV* is calculated as follows (Ross, 2007, pp. 38–39): if *X* is a discrete random variable with *k* = 1 to *K* outcomes (here: percentile impact classes) having a probability mass function *p*(*x*), then the *EV* of *X* is defined by

$$E(X) = \sum_{k=1}^K x \cdot p(x)$$

In other words, the *EV* of *X* is a weighted average of the possible values that *X* can take on, whereas each value being weighted by the probability – or relative frequency as proxy – that *X* assumes that value.

Table 1 presents the *EV*s for the three research groups. For a research group the numbering of the percentile impact classes (see the second column in the table) was multiplied with the relative frequencies (*p*) of papers, before the sum – the *EV* – over the products was calculated. For instance, the *EV* for research group 1 is the sum of the following products:  $1 \cdot 0.3 + 2 \cdot 0.1 + 3 \cdot 0.2 + 4 \cdot 0.1 + 5 \cdot 0.2 + 6 \cdot 0.1 = 3.1$ . As we operate with six percentile impact classes for the fictitious example, there is a maximal possible citation performance with an *EV* of 6 (all papers belong to the top 1%:  $6 \times 1$ ) and a minimal possible citation performance with an *EV* of 1 (all papers belong to the bottom 50%:  $1 \times 1$ ). The *EV* of each research group can be compared to a reference *EV* (for a medium performance) which can be obtained by the sum of the products of percentile class proportions with the numbering of classes:  $0.50 \cdot 1 + 0.25 \cdot 2 + 0.15 \cdot 3 + 0.05 \cdot 4 + 0.04 \cdot 5 + 0.01 \cdot 6 = 1.9$  (<50%, 50th, 75th, 90th, 95th, 99th). As the performance of research groups with *EV*s at about 1.9 is on medium-level, the three research groups in Table 1 with values between 3.0 and 3.9 performs better than a reference group. In agreement to the already presented

results, with an *EV* of 3.9 research group 2 is characterized by a higher citation performance than research group 1 (3.1) and research group 3 (3.0). Research group 2 comes much closer to the maximal possible *EV* of 6 than the other two research groups.

With the calculation of *EVs* on the base of percentile impact classes the proposals of Bornmann (2010) and Opthof and Leydesdorff (2010) for an optimized citation performance measurement of research groups were enhanced by a further metric. We are sure that this metric is an important but not the final step towards an optimal method to measure and compare citation performance.

## References

- Agresti, A. (2002). *Categorical data analysis*. Hoboken, NJ, USA: John Wiley & Sons, Inc.
- Anon (2010). Eleventh International Conference on Science and Technology Indicators. Book of Abstracts. Centre for Science and Technology Studies, Leiden, the Netherlands.
- Bornmann, L. (2010). Towards an ideal method of measuring research performance: Some comments to the Opthof and Leydesdorff (2010) paper. *Journal of Informetrics*, 4(3), 441–443.
- Bornmann, L., de Moya-Anegón, F., & Leydesdorff, L. (2010). Do scientific advancements lean on the shoulders of giants? A bibliometric investigation of the Ortega hypothesis. *PLoS One*, 5(10), e11344.
- 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.
- Boyack, K. W. (2004). Mapping knowledge domains: characterizing PNAS. *Proceedings of the National Academy of Sciences of the United States of America*, 101, 5192–5199.
- 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 Research. (2008). *Citation statistics. A report from the International Mathematical Union (IMU) in cooperation with the International Council of Industrial and Applied Mathematics (ICIAM) and the Institute of Mathematical Statistics (IMS)*. Berlin, Germany: International Mathematical Union (IMU).
- 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.
- Leydesdorff, L., & Opthof, T. (in press). Remaining problems with the “New Crown Indicator” (MNCS) of the CWTS. *Journal of Informetrics*.
- National Science Board. (2010). *Science and engineering indicators 2010, appendix tables*. Arlington, VA, USA: National Science Foundation (NSB 10-01).
- 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.
- Ross, S. M. (2007). *Introduction to probability models*. London, UK: Elsevier.
- van Raan, A. F. J. (2005). Measurement of central aspects of scientific research: Performance, interdisciplinarity, structure. *Measurement*, 3(1), 1–19.
- 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 Opthof and Leydesdorff. *Journal of Informetrics*, 4, 431–435.
- Waltman, L., van Eck, N. J., van Leeuwen, T. N., Visser, M. S., & van Raan, A. F. J. (2011). Towards a new crown indicator: Some theoretical considerations. *Journal of Informetrics*, 5, 37–47.

Lutz Bornmann\*

Max Planck Society, Office of Research Analysis and Foresight, Hofgartenstr. 8, D-80539 Munich, Germany

Rüdiger Mutz

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

\* Corresponding author.

E-mail address: bornmann@gv.mpg.de (L. Bornmann)

27 October 2010