Online Information Review

A simple method for excluding self-citation from the h-index: the b-index

Richard J.C. Brown,

Article information:

To cite this document:
Richard J.C. Brown, (2009) "A simple method for excluding self-citation from the h-index: the b-index", Online Information Review, Vol. 33 Issue: 6, pp.1129-1136, https://doi.org/10.1108/14684520911011043
Permanent link to this document:
https://doi.org/10.1108/14684520911011043

Users who downloaded this article also downloaded:

For Authors

About Emerald www.emeraldinsight.com

The current issue and full text archive of this journal is available at
www.emeraldinsight.com/1468-4527.htm

A simple method for excluding self-citation from the *h*-index: the *b*-index

Richard J.C. Brown
Analytical Science Team, National Physical Laboratory, Teddington, UK

Refereed article received
7 January 2009
Approved for publication
10 July 2009

Abstract

Purpose – The purpose of this conceptual paper is to present a simple, novel method for excluding self-citation from *h*-index values – the *b*-index.

Design/methodology/approach – The work described assumes that relative self-citation rate is constant across an author's publications and that the citation profile of a set of papers follows a Zipfian distribution, and from this derives a simple mathematical expression for excluding self-citation from *h*-index values.

Findings – It is shown that this new index is simply equal to the integer value of the author's external citation rate (non-self-citations) to the power three quarters, multiplied by their *h*-index. This value, called the *b*-index, does not require an extensive analysis of the self-citation rates of individual papers to produce, and appropriately shows the biggest numerical decreases, as compared to the corresponding *h*-index, for very high self-citers.

Practical implications – The method presented allows the user to assess quickly and simply the effects of self-citation on an author's *h*-index.

Originality/value – This paper provides a simple and novel method for excluding self-citation from the *h*-index and should be of interest to those interested in bibliometrics and databases of scientific literature.

Keywords Referencing, Serials, Publications, Peer-review

Paper type Conceptual paper

Introduction

The *h*-index has been proposed previously as a measure of scientific publishing esteem (Hirsch, 2005). The *h*-index of an individual author is defined as the maximum number of papers *h* by a scientist where each of those papers has received *h* or more citations. The *h*-index attempts to measure both the overall productivity of a scientist and the impact of this output, and addresses the weaknesses of traditional bibliometric indicators, such as the total number of publications and citations. It is recognised that the total number of publications does not provide a measure of the quality of these publications, while the total number of citations may be biased by relatively few, very highly cited works.

The strengths and weaknesses of the *h*-index have been discussed at length. The advantages of the *h*-index include that it is mathematically simple, it may be applied to any level of aggregation, it is a robust indicator (Rousseau, 2007), and it improves the quality of published work. Some criticisms of the *h*-index have included that it is

This paper expresses the personal opinions of the author and does not necessarily represent the views of the National Physical Laboratory or NPL Management Limited.

Online Information Review
Vol. 33 No. 6, 2009
pp. 1129-1136
© Emerald Group Publishing Limited
1468-4527
DOI 10.1108/14684520911011043

bounded by the total number of publications, it does not recognise the context of the citations received, it is limited by the completeness of the database used to do the calculation, and it does not account for the number of authors on a paper. Indeed other similar measures such as the *g*-index (which is the maximum number of papers that have each received at least g^2 or more citations (Egghe, 2006)), the *AR*-index (which takes into account the age of a publication and the intensity of citation of the *h*-index publications (Jin *et al.*, 2007)) and the *w*-index (which is the maximum number of papers that have each received at least $10w$ or more citations (Wu, 2008)), to name only a few, have been proposed as improvements.

Interestingly, Hirsch concluded that self-citation would not significantly affect the *h*-index of an individual author. Although self-citations can obviously increase a scientist's *h*-index, he contended that their effect on the *h*-index is much smaller than on the total citation count. This is an assumption that has been challenged by many studies (Kelly and Jennions, 2006; Schreiber, 2007) that have shown average decreases in the *h*-index of 12 per cent with self-citations excluded. The crux of the argument is not whether self-citation affects the *h*-index, but the extent to which it does (Zhivotovsky and Krutovsky, 2008).

Of course it is not unreasonable that self-citation should occur when an author refers to relevant previous work or avoids repetition of previous experimental set-ups. Indeed the self-citation percentage may be a high percentage of all citations if much of the author's previous work is relevant and few other sources are cited. Self-citation has been shown to vary by discipline and by origin of publication, but average self-citation rates between 10 and 40 per cent are common (Thuis and Glanzel, 2006). This notwithstanding, self-citation is a tool entirely within the author's power to use and may be open to abuse to increase total citation count and, through very targeted citation, to specifically increase the *h*-index. It seems reasonable therefore that the *h*-index be considered without self-citations included for a more equitable comparison of authors in the same field – and at the very least as an additional piece of bibliometric information.

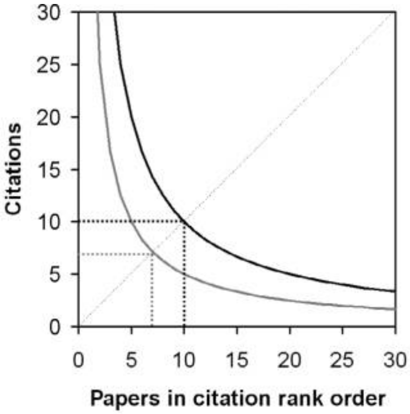
While it is easy to use the ISI Web of Science database (Institute for Scientific Information, 2009) to determine the *h*-index (Jacsò, 2008), it is not currently a simple task to determine the *h*-index using only external citations. External citations are defined as non self-citations. Rigorously, this would involve determining the external citations for each paper by each author and then re-ranking these to determine a revised *h*-index (Kelly and Jennions, 2006) – a very time-consuming task. (Perhaps in the future ISI Web of Science will offer this functionality?). Indeed no other proposed bibliometric indices have addressed the removal of self-citation from the *h*-index satisfactorily. Therefore this paper offers a simple and quick solution to correct the *h*-index to account for the overall self-citation rate of an author based on a simple mathematical description of the properties of the *h*-index (a more thorough mathematic description of the properties of the *h*-index is available in Glanzel (2006)). The resulting indicator is referred to as the *b*-index.

Methods

The method presented in this paper makes two major assumptions. The first of these is that relative self-citation rates are uniform across all publications of an individual author. That is to say that if an author has an overall self-citation rate of 50 per cent, each individual paper has a self-citation rate of 50 per cent. To a first approximation this seems like a reasonable assumption and has some support in practice by those examining self-citation rates for individual papers (Kelly and Jennions, 2006; Schreiber, 2007, 2008). While studies have shown that the percentage of self-citations is seen to decrease as the number of overall citations increases (Aksnes, 2003), the change in relative self-citation rate was only observed to be 5 percentage points per every 20 additional citations. A diagrammatic representation of the *h*-index for an author with and without self-citation included, assuming a constant relative self-citation rate, is displayed in Figure 1.

In fact the assumption made above is more general than stated. It is not crucial to the process that the relative self-citation rate is constant across all publications, but more so that when self-citations are removed and the publications re-ordered according to citation rank, the new profile is the same as the old one, but a sub-multiple of it (as seen in Figure 1). The accuracy of this approximation notwithstanding, this is the assumption that will be used in defining the method to exclude self-citation described herein. It is additionally important to recognise that the external citation rate for an author is not currently easy to determine, as this is a functionality that ISI Web of Science does not currently provide. If one assumes that the number of citations received per paper is approximately equal for external citations and self-citations then the external citation rate may be estimated as the ratio of the number of "citing articles without self-citations" to the total number of "citing articles" supplied on the ISI Web of Science's Citation Report function. However this approximation does not affect the conceptual method described in this paper.

The second assumption made as part of this method is that the profile of citations for an individual author follows a Zipfian distribution (Brown, 2007). It has been shown



Note: Self-citations are included (solid black line) and self-citations excluded (solid grey line). The h-index of an imaginary author is calculated with self-citations included (dashed black lines) and self-citation excluded (dashed grey lines). The self-citation rate of 50 per cent has been assumed either a) to apply equally across all papers, or b) once the self-citations are removed and the papers re-ranked the profile is assumed to be the same as, but a multiple of 0.5 of the original

Figure 1.
Diagrammatic representation of the citations for a set of papers in citation rank order

previously that the profile of citations for a collection of papers by many authors follows a Zipfian distribution (Newman, 2005; Gupta *et al.*, 2005) and since such a distribution is the result of the sum of distributions for all the authors in a set, it is not unreasonable to suggest that the average distribution for an individual author will also be Zipfian. Thus it may be written that:

$$f = \frac{A}{r^n} \tag{1}$$

where f is the total citation count (the cumulative number of citations received) of the rth ranked paper (ranked by total citations in decreasing order), n is an exponent characterising the Zipfian distribution (see Brown (2007) for more details) and A is the total citation count for the most highly cited paper ($r = 1$). The h-index is defined as the point when the rank is numerically equal to the total citation count of the paper at that rank, $r = f$, which yields:

$$r = \frac{A}{r^n} \tag{2}$$

and thus;

$$r^{n+1} = A \tag{3}$$

Let us now imagine that the individual author in question has a fractional external citation rate k ($0 \leq k \leq 1$) across all papers, such that with self-citations excluded we may rewrite equation (1) as:

$$f_g = \frac{Ak}{r_g^n} \tag{4}$$

where $f_g = kf$, and r_g is the ranking of papers by total citations excluding self-citation. At the new h-index point, $r_g = f_g$, which when substituted into equation (4) yields:

$$r_g^{n+1} = Ak \tag{5}$$

The ratio, R, of the h-index not including self-citations to that including self-citations, $R = \left(r_g/r\right)$, can then be derived by dividing equation (5) by equation (3):

$$\frac{r_g^{n+1}}{r^{n+1}} = k \tag{6}$$

and thus

$$R = k^{1/(n+1)} \tag{7}$$

Therefore to obtain the h-index of an author without self-citations using this method, which will now be referred to as the b-index, we may simply multiply the h-index, obtained including all citations, by R.

Results and discussion

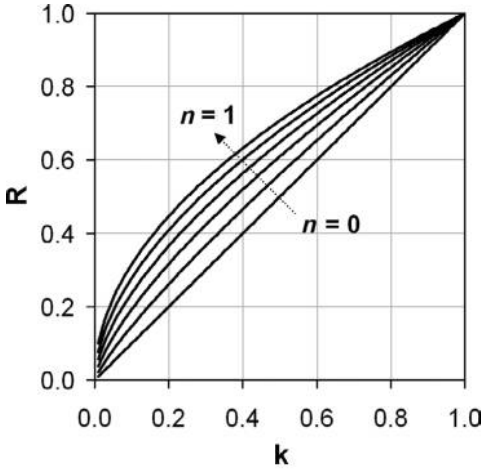
The derivation above demonstrates that the difference between the *h*-index and the *b*-index depends only on an author's external citation rate, and on the exponent in the Zipfian relationship.

The dependence of R on external citation rate, k, and Zipfian exponent, n, is displayed in Figure 2, which shows that when $n = 0$, the case when the citation count is constant for all papers, the *h*-index decreases linearly as a function of decreasing external citation rate. As n increases the decrease in the *h*-index is less pronounced as the external citation count decreases. Hence for $n = 1$, an author with an external citation rate of 0.5 would only expect the numerical value of their *b*-index compared to their *h*-index to decrease by a factor of 0.7 on the exclusion of self-citations. Previous studies have shown that the fit to a Zipfian profile for citations on the ISI Web of Science database (excluding very highly cited papers) was best for $n = 1/3$ (Gupta *et al.*, 2005; Redner, 1998; Tsallisa and de Albuquerque, 2000) which yields from equation (7):

$$R = k^{3/4} \tag{8}$$

In this case, for an external citation rate of 0.75, when self-citations are excluded the *h*-index would be expected to decrease by a factor of 0.8 to yield the corresponding *b*-index. For an external citation rate of 0.5, when self-citations are excluded the *h*-index would be expected to decrease by a factor of 0.6 to yield the corresponding *b*-index. In all cases where self-citation is present the calculated *b*-index yields a decrease in the corresponding *h*-index by at least 1, as the *b*-index has been rounded down to the nearest integer to represent the greatest integer number, *b*, of papers with *b* external citations – parts of papers are meaningless in this context. Hence the relationship between the *h*-index and the corresponding *b*-index may be described as:

$$b = \text{int}\left(hk^{3/4}\right) \tag{9}$$



Note: $n = 0$ (bottom line), 0.2, 0.4, 0.6, 0.8, and 1 (top line)

Figure 2.
The relationship between the fractional change in the *h*-index upon excluding self-citations, *R*, and the external citation rate, *k*, for various exponentials characterising the Zipfian distribution of citations across a set of publications

Where int (x) is the integer value of x, and b and h are the values of the b-index and h-index respectively. Therefore, the b-index is initially very sensitive to self-citation – if any is present the calculated value of the b-index will be less than the h-index. Aside from this property, Figure 2 shows that the b-index displays a sensitivity to self-citation that increases rapidly as the self-citation rate increases. This effect is displayed in Figure 3 for an author with an h-index of 10. A brief trial of this new technique on real publication data from the ISI Web of Science database produced a similar reduction in the numerical values of h-indices of about 15 per cent, as compared with studies that had systematically removed self-citations from individual publications (Kelly and Jennions, 2006; Schreiber, 2007, 2008), but the data was produced in only a fraction of the time.

As a further example the h-index and the b-index of the author of this paper were calculated. When the analysis was performed using the ISI Web of Science database, the author was listed as having 52 publications with 400 total citations and an h-index of 10. The external citation rate was estimated as the ratio of the total number of citing articles excluding self-citation to the total number of citing articles – both values being available from the Citation Report function of the Web of Science database. This yielded a ratio of 262 to 312 which when expressed as a decimal is 0.84. The author's b-index is then given, by equation (9), as the integer value of 10 multiplied by 0.84 to the power three-quarters. This yields a b-index value of 8, a reduction in the corresponding numerical h-index value of 20 percent.

Conclusion

A novel and simple way of correcting h-index values to remove self-citations has been proposed. This is referred to as the b-index. This method avoids the laborious process of removing self-citations from individual publications. The b-index for an author is simply equal to the integer value of the author's external citation rate to the power three quarters, multiplied by their h-index. While this method provides a useful and efficient shortcut, it is hoped in the future that the ISI Web of Science database will

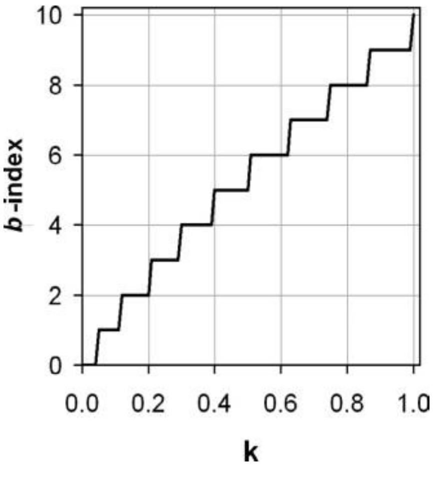


Figure 3.
The relationship between the calculated b-index for an author with an h-index of 10, and their external citation rate, k

provide *h*-indices without self-citations as an extra functionality – the Scopus database (Scopus, 2009) already offers an *h*-index without self-citations.

References

Aksnes, D.W. (2003), “A macro study of self-citation”, *Scientometrics*, Vol. 56 No. 2, pp. 235-46.

Brown, R.J.C. (2007), “The use of Zipf’s law in the screening of analytical data: a step beyond Benford”, *Analyst*, Vol. 132 No. 4, pp. 344-9.

Egghe, L. (2006), “An improvement of the *h*-index: the *g*-index”, *International Society for Scientometrics and Infometrics Newsletter*, Vol. 2 No. 1, pp. 8-9.

Glanzel, W. (2006), “On the h-index – a mathematical approach to a new measure of publication activity and citation impact”, *Scientometrics*, Vol. 67 No. 2, pp. 315-21.

Gupta, H.M., Campanha, J.R. and Pesce, R.A.G. (2005), “Power-law distributions for the citation index of scientific publications and scientists”, *Brazilian Journal of Physics*, Vol. 35 No. 4A, pp. 981-6.

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*, Vol. 102 No. 46, pp. 16569-72.

Institute for Scientific Information (2009), “Web of knowledge”, (online), available at: http://isiknowledge.com (accessed 15 January 2009).

Jacsò, P. (2008), “The pros and cons of computing the *h*-index using Web of Science”, *Online Information Review*, Vol. 32 No. 3, pp. 673-88.

Jin, B.H., Liang, L.M., Rousseau, R. and Egghe, L. (2007), “The R- and AR-indices: complementing the *h*-index”, *Chinese Science Bulletin*, Vol. 52 No. 6, pp. 855-63.

Kelly, C.D. and Jennions, M.D. (2006), “The h-index and career assessment by numbers”, *TRENDS in Ecology and Evolution*, Vol. 21 No. 4, pp. 167-70.

Newman, M.E.J. (2005), “Power laws, Pareto distributions and Zipf’s law”, *Contemporary Physics*, Vol. 46 No. 5, pp. 323-51.

Redner, S. (1998), “How popular is your paper? An empirical study of the citation distribution”, *European Physical Journal B*, Vol. 4 No. 2, pp. 131-4.

Rousseau, R. (2007), “The influence of missing publications on the Hirsch index”, *Journal of Infometrics*, Vol. 1 No. 1, pp. 2-7.

Schreiber, M. (2007), “Self-citation corrections for the Hirsch index”, *Europhysics Letters*, Vol. 78, Art No. 30002, 6 pp..

Schreiber, M. (2008), “The influence of self-citation corrections on Egghe’s g-index”, *Scientometrics*, Vol. 76 No. 1, pp. 187-200.

Scopus, Elsevier B.V. (2009), (online), available at: www.scopus.com (accessed 1 July 2009).

Thuis, B. and Glanzel, W. (2006), “The influence of author self-citations on bibliometric meso-indicators. The case of European universities”, *Scientometrics*, Vol. 66 No. 1, pp. 71-80.

Tsallisa, C. and de Albuquerque, M.P. (2000), “Are citations of scientific papers a case of nonextensivity?”, *European Physical Journal B*, Vol. 13 No. 4, pp. 777-80.

Wu, Q. (2008), “The w-index: a significant improvement on the *h*-index”, (online), available at: http://arxiv.org/abs/0805.4650 (accessed 15 January 2009).

Zhivotovsky, L.A. and Krutovsky, K.V. (2008), “Self-citation can inflate *h*-index”, *Scientometrics*, Vol. 77 No. 2, pp. 373-5.

About the author

Richard J.C. Brown is a principal research scientist in the analytical science team at the National Physical Laboratory. In 2008 he was awarded the 34th SAC Silver Medal by the Royal Society of Chemistry for his contributions to environmental analytical chemistry and electroanalytical chemistry, and the CITAC Award for the most important paper in metrology in chemistry for his work on mercury vapour measurement. He has published more than 50 peer-reviewed papers and is a Fellow of the Royal Society of Chemistry. Richard J.C. Brown can be contacted at: richard.brown@npl.co.uk

This article has been cited by:

1. BIBLIOGRAPHY 407-484. [CrossRef]
2. Tariq Ahmad Shah Department of Library and Information Science, University of Kashmir, Srinagar, India Sumeer Gul Department of Library and Information Science, University of Kashmir, Srinagar, India Ramesh C Gaur Central Library, Jawaharlal Nehru University, New Delhi, India . 2015. Authors self-citation behaviour in the field of Library and Information Science. *Aslib Journal of Information Management* **67**:4, 458-468. [Abstract] [Full Text] [PDF]
3. Lorna Wildgaard, Jesper W. Schneider, Birger Larsen. 2014. A review of the characteristics of 108 author-level bibliometric indicators. *Scientometrics* **101**:1, 125-158. [CrossRef]
4. Emilio Ferrara, Alfonso E. Romero. 2013. Scientific impact evaluation and the effect of self-citations: Mitigating the bias by discounting the h-index. *Journal of the American Society for Information Science and Technology* **64**:11, 2332-2339. [CrossRef]
5. Colleen J. Goode, Lauren B. McCarty, Regina M. Fink, Kathleen S. Oman, MaryBeth Flynn Makic, Mary E. Krugman, Lisa Traditi. 2013. Mapping the Organization. *JONA: The Journal of Nursing Administration* **43**:9, 481-487. [CrossRef]
6. Lin Zhang, Bart Thijs, Wolfgang Glänzel. 2011. The diffusion of H-related literature. *Journal of Informetrics* **5**:4, 583-593. [CrossRef]