SEPT 12 – 16

2011

VIENNA

h-index and Related Measures – Part 2:

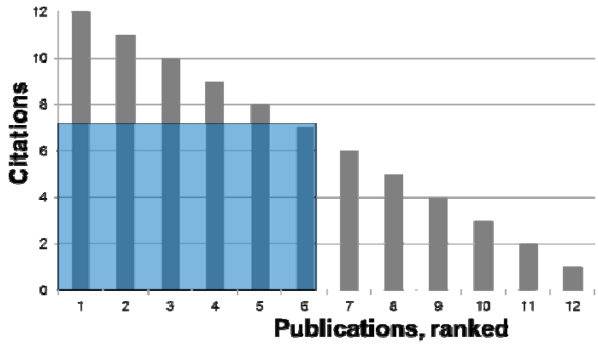
Variants and Extension of the h-index

european
summer school
for scientometrics

Variants and Extension of the h-index

Jeffrey Demaine

Citations covered by the h-index



Revisiting the concept of the h-index as expressed in the form of a graph, the most highly-cited articles are to the left and the less-cited articles are collected at the right of the frequency distribution. The *h-index* accounts for all the citations within the blue square. This set of publications (which could represent the output of a single author or of a whole institution) has an *h-index* of 6 because the 6th-most cited article has been cited at least six times (in this case, seven). From the graph we can see that there are many citations both above and to the right of the blue square that are not accounted for by the *h-index*. For example, the 7th publication has only been cited 6 times, so does not contribute to the *h-index*.

Similarly, citations located to the upper left of the graph belong to publications that have been cited more than *h* times. These citations over and above the value of *h* are ignored. As will be discussed later, many variations of the *h-index* have been proposed to take these citations into account.

Calculating the h-index

Practically speaking, how does one calculate the *h-index*? Once a set of publications is identified, their bibliographic metadata, including citations of each article, is collected. One might then use a spreadsheet such as Excel to sort the records by the number of citations received. The records are put in order of their decreasing citation counts and a rank is then assigned to each record, starting with the most frequently cited (i.e. the one with the highest citation count). For example, *Table 1* shows data pertaining to a set of twelve publications, ranked by the number of times each has been cited. To find the *h-index*, read down the *Times Cited* column until the value matches (or is less than) the rank value. In this example, the author has an *h-index* of ten.

Rank	Times cited
1	45
2	38
3	23
4	21
5	17
6	17
7	15
8	13
9	11
10	**10**
11	4
12	1

Table 1. Ranked citations

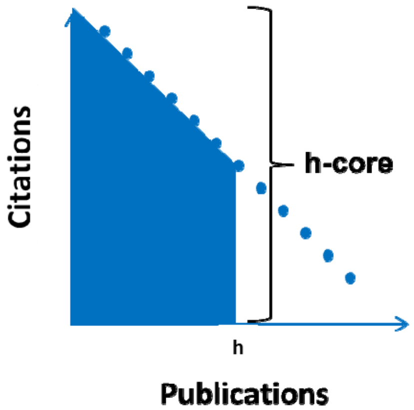
Note that while the first nine articles have received more than ten citations each, it is only the first ten of those citations that count towards this author's h-index. So the number of citations that are accounted for by this researcher's *h-index* is 10 x 10. Indeed, the number of citations covered by an h-index is simply the square of h.

Citations missed by the h-index

One of the criticisms of the *h-index* is that two researchers may both have the same *h-index* yet have very different citation profiles. A top researcher would appear to be equal to a less successful one because citations that exceed the level of h are ignored. Many of the variants of the *h-index* are designed to measure the differences between these two kinds of distributions.

The Hirsch core

In examining the properties of the *h-index*, Rousseau (2006) coined the term "Hirsch core" to refer to all of the citations received by the first "*h*" ranked articles. Although the *Hirsch core* (alternatively called the "h-core") was not proposed as an alternative to the *h-index*, it is a convenient way of referring to the overall impact that a researcher's best articles have had (including those covered by h-index).



Variants and Extensions of the h-index

A number of variants of the *h-index* have been proposed. Each tries to compensate for one or more shortcomings of the *h-index* (or one of its variants). The most widely-known variant of the *h-index* is perhaps the *g-index*. It was proposed by Leo Egghe (2006) as a way to capture more of the citations that fall outside the coverage of the *h-index*. It is determined by comparing the *cumulative total* of citations in a ranked list of publications against the rank. When this total exceeds (or equals) the *square* of the rank, the g-index has been found.

An easier-to-understand definition is given by Quesada (2010): *"The g-index is the maximum number g of papers by r such that the average number of citations of the g papers is at least g."* An easy way to determine the g-index is by calculating the h-index of the average citation count:

$$g = h(a^x)$$

Where a^x is the vector of average citations.

By counting the average citations in the h-core, the *g-index* captures more of the impact of those highly-cited publications whose impact the h-index leaves out because they exceed h citations. It is therefore a variation of the h-index.

Quesada showed that for both the h and g-indexes, it becomes harder to increase either index as publication list grows. There is a *"decreasing marginal contribution of citations and papers: the larger the output, the smaller the impact on the index caused by an additional paper or an additional citation."*

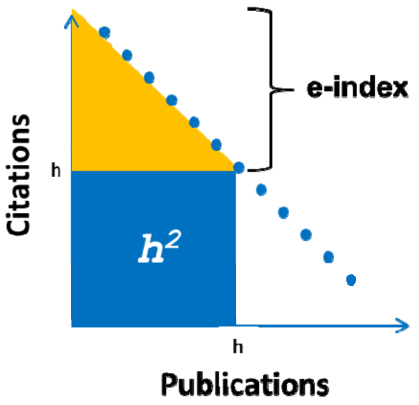
The following year Egghe was again involved in an influential paper: Jin *et al* (2007) proposed three new indicators: the A, R, and AR-indexes. The *A-index* is the average number of citations of articles in the h-core. However averages are susceptible to extreme values and so it overly influenced by one or two "hit" papers, and may not reflect the true impact of a researcher. The *R-index* (R stands for "root") compensates for this by taking the *square-root* of the sum of citations in the h-core. The R-index thus measures the citation intensity of the h-core. It is much less influenced by outliers than the *A-index*. The *AR-index* is a variation of the R-index which takes the age of articles into consideration, dividing the number of citations received by an article by the years since the publication of the article. Interestingly, the AR value can decrease over time. This makes it a more accurate measure of the current status of a researcher's career than the h-index which can only increase, even if the researcher has long been retired.

There are many more variants of the *h-index* than those just discussed. In addition, there are metrics that extend the h-index by taking factors other than citations into consideration as well. These are namely the change in the rate of citations over time, and the relative importance of authorship and/or citations as a function of the number of authors contributing to (or the number of citations received by) a given article. These extensions of the h-index bring another dimension to the evaluation of scientific productivity. Yet while many deserve to be used in the management of research, none seem to have caught on outside the bibliometrics community.

Some variants and extensions of note include:

- **hg**-index *(Alonso et al., 2009)*
- $\mathbf{h_{\alpha}}$ -index *(Eck & Waltman, 2008)*
- **normalized, contemporary, trend h**-index (*Sidiropoulos et al., 2007*)
- **Tapered h**-index (*Anderson et al., 2008*)
- **Rational h**-index (*Ruane & Toi, 2008*)
- **Dynamic h-type** (*Rousseau & Ye, 2008*)
- $\mathbf{h_m}$ -index (*Schreiber, 2008*)
- $\mathbf{h_I}$ -index (*Batista et al., 2006*)
- *and others...*

The e-index



A recent addition to the list of variants is the *e-index* (Zhang, 2008), which is calculated by simply subtracting the citations in the *h-index* from all those in the *h-core* (remember that the number of citations that the *h-index* accounts for is simply the square of *h*). This simple step captures all of the impact of highly-cited papers, and is therefore a convenient metric that can be used to complement the h-index and not necessarily to replace it.

Recommendations on the use of the h-index

Now that the concept behind the *h-index* has been discussed in detail, a look at some examples of where this bibliometric indicator is being used. Issues of data provenance and intent illustrate the ways in which the *h-index* is used and mis-used.

General shortcomings of bibliometric data

As with all bibliometric measures, the characteristics of the research field under study or the shortcomings of the data used may have a negative effect on calculations of the h-index. For example, articles published in languages other than English may not be indexed by citation databases. They would therefore be difficult to include in any calculation of an *h-index*. Similarly, some conference proceedings may also not be counted or poorly covered. For a field such as computer science in which much communication is through conferences, it may be difficult to calculate a valid *h-index*.

A particular problem is that of *author homonyms*. Some names are very common and an inflated value of h may be arrived at if one does not filter out articles from authors who share a name with the person being studied. Note that these problems are not particular to the *h-index*, but affect many bibliometric calculations.

Data sources for bibliometric indicators

There are only three generally available sources of citation data used in calculating the *h-index*. The *Web of Science* and *Scopus* are reasonably comparable in terms of their accuracy and coverage. However, this does not mean that *h-index* from one is directly comparable to an *h-index* from the other. Differences in coverage (particularly of conferences) mean that a bibliometric indicator generated with data from one will be similar to, but not necessarily the same as, the same indicator generated with data from the other. A third source of data, *Google Scholar*, is also popular because it is free and can be queried automatically by other applications. It covers many online items that *Web of Science* and *Scopus* do not, but includes many duplicates and spurious records. A great deal of post-processing is required when using *Google Scholar*, as not all of the millions of “hits” it finds are likely to be valid (Alonso *et al.*, 2009).

Usage of the h-index

The ease of calculating the *h-index*, coupled with people's affinity for numerical indicators of value, has led to the appearance of the *h-index* in a variety of contexts. For example, the following bibliographic citation from a researcher's list of publications is complemented by three metrics of the prestige of the journal (in bold) in which the article was published:

> Byrne, J., Kendrick, M., and Sroaf, D. 2007 "The Park Made of Oil: A Historical Political Ecology of Park Development in the Baldwin Hills" *Local Environment* Vol. 12(2) pp. 153-181. **[ISI=n/a; H index=15, SJR= 0,044]** cited 1 time

Here, the researcher has determined that this journal has an *h-index* of 15. Because h-indexes are typically used in the context of individuals, a cursory look at these numbers might lead one to believe that it is the *author* who has an h-index of 15, which is not the case. Whatever the author's success in their field, note that the article in question has only been cited once. So while the journal may be reasonably "impactful", the article itself is not. The lesson here is that when one encounters an *h-index*, it is important to stop and consider why it is being used and what it actually refers to (in this case, the *journal h-index*).

Chemistry World H-index ranking of living chemists

The journal *Chemistry World* ranked chemists by their h-indexes and published the table as a special feature in the journal. This is a good example of a valid use of this bibliometric indicator: All the researchers are from the same field, making a comparison of their h-scores reasonable. Moreover, the journal plays a neutral role in this field and does not seek to promote one individual or institution over another. It is simply using the h-index to describe the field in a novel way.

Rank	Name	h-index	Field
1	Whitesides, G. M.	155	Organic
2	Karplus, M.	139	Theoretical
3	Corey, E. J.#	138	Organic
4	Heeger, A. J.#	128	Organic
5	Huber, R.#	122	Bio
6	Wüthrich, K.#	120	Bio
7	Bax, A.	118	Bio
8	Schleyer, P. v.	117	Organic
9	Lehn, J. M. #	114	Organic
10	Bard, A. J.	113	Analytical
10	Gratzel, M.	113	Physical
10	Hoffmann, R.#	113	Theoretical
13	Schreiber, S. L.	112	Bio
14	Scheraga, H. A.	111	Bio
15	Fersht, A. R.	105	Bio
15	Frechet, J. M.	105	Inorganic
15	Truhlar, D. G.	105	Theoretical
18	Marks, T. J.	104	Inorganic
18	Trost, B. M.	104	Organic
20	Gray, H. B.	103	Inorganic

Interestingly, those names followed by a *pound symbol* (#) indicate those chemists who have been awarded the Nobel Prize. There would seem to be a good correlation between a chemist's *h-index* and the recognition they have received. Thus, for all its faults, the *h-index* is a good reflection of the status of chemists.

Scholarometer

The ease with which bibliographic data can be mined from *Google Scholar* has led to the appearance of the *h-index* in various *mash-ups*. The *Publish-or-Perish* application is perhaps the most widely known. Another example is *Scholarometer*, a browser plugin that generates interactive social network graphs based on the citations in a field. When the user holds the mouse over a vertex in the network, a tool-tip window appears listing various bibliometric indicators for that person, including the h-index and the g-index. While the graphs are eye-catching and provide a description of the field that goes beyond straightforward publication lists, remember that the data behind the graph comes from *Google Scholar* and will therefore not be comparable to bibliometric indicators derived from other databases.

Recommendations

Bearing these issues in mind, and remembering the shortcomings inherent in all bibliometric data, one is wise to consider that the *h-index* cannot simply be taken at face value. One must understand where the data came from, and must be aware of the context from which it is derived. It is therefore problematic to use the *h-index* for comparisons between researchers from different fields, with a different amount of research experience, or from countries with different languages.

Many research articles evaluating the precision of the *h-index* (and its variants) conclude that one indicator alone is not sufficient to express the quantity and quality of a researcher's output. For example Bornmann, Mutz, and Daniel (2009) performed a factor-analysis of nine different h-type indices. They found that the indexes could be divided into two groups: those that captured the *quantity* of a researcher's output (the h-index, g-index, h^2-index, and the m-quotient), and those that captured the *quality* of a researcher's publications (the a, m, r, ar, and h_w-indexes). They recommend therefore that a pair of indicators (one from each group) be used to evaluate the overall impact of a researcher.

Type	Index name	Author(s)	Notes
	h-index	Hirsch (2005)	Simple to calculate, objectively derived from popular DBs, robust to poorly-cited 'tail'. But ignores impact of important papers cited more frequently than h times.
Variants	g	Egghe (2006)	Gives more weight to highly-cited papers but is not robust to influence of outliers.
Variants	hg	Alonso *et al*. (2009)	$\sqrt{h \cdot g}$
Variants	h_α	Eck & Waltman (2008)	A generalized form of the h-index.
Variants	*A*	Jin (2006)	$A = \frac{1}{h}\sum_{j=1}^{h} cit_j$
Variants	*R*	Jin (2007)	$R = \sqrt{\sum_{j=1}^{h} cit_j}$
Variants	m	Bornman *et al*. (2008)	
Variants	$h^{(2)}$	Kosmulski (2006)	
Variants	e	Zhang (2009)	$e^2 = \sum_{j=1}^{h} cit_j - h^2$ Independent yet complementary to the h-index. Useful for evaluating highly-cited researchers or differentiating between researchers with the same h-index.
Variants	normalized h	Sidiropoulos *et al*. (2007)	$h^n = \frac{h}{N_p}$
Variants	tapered h	Anderson *et al*. (2008)	Takes all citations into account.
Variants	rational h	Ruane & Toi (2008)	$h_{rat} = (h+1) - \frac{n_c}{2 \cdot h + 1}$ Increases in smaller steps than h: has more granularity.
Extensions	h_w	Egghe & Rousseau (2008)	The h-value weighted by citation impact.
Extensions – temporal	m-quotient	Hirsch (2005)	$m = \frac{h}{y}$
Extensions – temporal	contemporary h	Sidiropoulos *et al*. (2007)	Older articles have less 'weight'. Identifies promising new researchers.
Extensions – temporal	trend h	Sidiropoulos *et al*. (2007)	Emphasizes recent citations, identifying researchers who are 'hot' now, even if their articles were are old.
Extensions – temporal	dynamic h-type	Rousseau & Ye (2008)	Tries to differentiate static *vs.* increasing h-indexes
Extensions – individuals	h_I	Batista et al. (2006); Imperial & Rodriguez-Navarro (2007)	Divides the standard h-index by the average number of authors in the articles that contribute to the h-index, in order to reduce the effects of co-authorship.
Extensions – individuals	h_m	Schreiber (2008)	Uses fractional paper counts instead of reduced citation counts to account for shared authorship of papers. Determines the **multi-authored h_m index** based on the resulting effective rank of the papers using undiluted citations.
Extensions – individuals	fractional counting of papers	Egghe (2008)	Same as h_m-index

Table 2: Selected variants and extensions of the h-index

References

S. Alonso, F.J. Cabrerizo, E. Herrera-Viedma, F. Herrera (2009). h-Index: A review focused in its variants, computation and standardization for different scientific fields. *Journal of Informetrics* **3**: 273-289.

Anderson T.R., Hankin R.K.S., Killworth P.D. *(2008)* **Beyond the Durfee square: Enhancing the h-index to score total publication output**. *Scientometrics **76**(3) 577-588.* doi:10.1007/s11192-007-2071-2.

Batista P.D., Campiteli M.G., Kinouchi O., Martinez A.S. *(2006)* **Is it possible to compare researchers with different scientific interests?**. *Scientometrics **68**(1) 179-189.* doi:10.1007/s11192-006-0090-4.

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.* doi:10.1002/asi.21016.

Costas, R.; Bordons, M. (2007) The h-index: Advantages, limitations and its relation with other bibliometric indicators at the micro level. *Journal of Informetrics* **1**: 193-203.

van Eck N.J., Waltman L. *(2008)* **Generalizing the h- and g-indices**. *Journal of Informetrics **2**(4) 263-271.* doi:10.1016/j.joi.2008.09.004.

Egghe L., Rousseau R. *(2008)* **An h-index weighted by citation impact**. *Information Processing & Management 44 (2) 770-780.* doi:10.1016/j.ipm.2007.05.003

Hirsch J.E. (2005) **An index to quantify an individual's scientific research output**. *Proceedings of the National Academy of Sciences 102 (2005) 16569-16572.* doi:10.1073/pnas.0507655102.

Jin, B.; Liang, L.; Rousseau, R.; Egghe, L. (2007) The R- and AR-indices: Complementing the h-index. *Chinese Science Bulletin* **52**(6):855-863.

Quesada, Antonio (2010) Axiomatics for the Hirsch index and the Egghe index. *Journal of Informetrics* **5**(3) 476-480.

Rousseau, R. (2006) New developments related to the Hirsch index. *Science Focus,* **1**(4), p..23-25, 2006. (in Chinese). English version available at: E-LIS: code 6376 [preprint also available at http://sci2s.ugr.es/hindex/pdf/Hirsch_new_developments.pdf]

Rousseau R., Ye F.Y. (2008) **A proposal for a dynamic h-type index**. *Journal of the American Society for Information Science and Technology **59**(11) 1853-1855.* doi:10.1002/asi.20890.

Ruane F., Tol R.S.J. (2008) **Rational (successive) h-indices: An application to economics in the Republic of Ireland**. *Scientometrics **75**(2) 395-405.* doi:10.1007/s11192-007-1869-7.

Schreiber M. *(2008)* **To share the fame in a fair way, h(m) modifies h for multi-authored manuscripts**. *New Journal of Physics **10** 040201.* doi:10.1088/1367-2630/10/4/040201.

Sidiropoulos A., Katsaros D., Manolopoulos Y. *(2007)* **Generalized Hirsch h-index for disclosing latent facts in citation networks**. *Scientometrics **72**(2) 253-280.* doi:10.1007/s11192-007-1722-z.

Zhang C.T. *(2009)* **The e-Index, Complementing the h-Index for Excess Citations**. *PLoS ONE **4**(5):e5429.* doi:10.1371/journal.pone.0005429.