Reflections on recent developments of the h-index and h-type indices

Ronald Rousseau[1,2]

01 June 2008

Abstract

A review is given of recent developments related to the h-index and h-type indices. Advantages and disadvantages of the h-index are listed. Properties of generalizations of the h-index are shown. Woeginger's axiomatic approach to the study of impact indicators is highlighted as the most exciting development in recent months.

1 Introduction

Since its introduction by Hirsch in 2005 (Hirsch, 2005) the h-index has evolved into a real hype (Ball, 2007). In the Web of Science alone Hirsch's article has been cited already 129 times (May 30, 2008), and many more articles related to the h-index are in the pipeline. In this note we give an overview of some of the recent developments related to the h-index. Already published reviews on this topic include (Bornmann & Daniel, 2007; Bornmann et al., 2008).

Let us first recall the definition of the h-index. Consider the list of articles [co-] authored by scientist S, ranked according to the number of citations each of these articles has received. Articles with the same number of citations are given different rankings (for the moment the exact order does not matter). Then scientist S' Hirsch index is h if the first h articles received each at least h citations, while the article ranked h+1 received strictly less than h+1 citations. Stated otherwise: scientist S' Hirsch index is h if h is the largest natural number (representing a rank) such that the first h publications received each at least h citations. The first h articles in such a ranked list form the h-core. It seems reasonable to rank articles with the same number of citations in anti-chronological order, i.e. youngest article first. If necessary, further ties can be solved by taking the impact factor of the journal in which the article is published into account (lowest IF first). It is further noted that if the last article in the list occupies rank r and receives c > r citations then this scientist's h-index is equal to r.

2 Advantages and disadvantages of the h-index

We first present a list of advantages and then add comments. This list is mainly derived from Hirsch (2005), Glänzel (2006) and Liu & Rousseau (2007).

Advantages of the h-index

- It is a mathematically simple index.
- It is a better index than total number of publications or total number of citations alone.
- It encourages high quality (or at least highly visible) work.
- The h-index can be applied to any level of aggregation. This means: not only to scientists but also to different groups of scientists, such as all scientists belonging to a research institute, belonging to a university, even belonging to the same country.

[1] KHBO (Association K.U.Leuven), Industrial Sciences and Technology, Zeedijk 101, B-8400 Oostende, Belgium; E-mail: ronald.rousseau@khbo.be

[2] K.U.Leuven, Steunpunt O&O Indicatoren, Dekenstraat 2,B-3000 Leuven, Belgium

H. Kretschmer & F. Havemann (Eds.): Proceedings of WIS 2008, Berlin

Fourth International Conference on Webometrics, Informetrics and Scientometrics & Ninth COLLNET Meeting

Humboldt-Universität zu Berlin, Institute for Library and Information Science (IBI)

- It combines two types of activity (in the original setting this is citation impact and publications).
- Increasing the number of publications alone does not have an immediate effect on this index.
- It is a robust indicator in the sense that small errors in data collection have no or little effect.
- Single peaks (top publications) have hardly any influence on the h-index.
- In principle, any document type can be included.
- Publications that are hardly ever cited do not influence the h-index.

Comments: The first items in this list of advantages were already brought forward by Hirsch (2005). The fact that the h-index is a robust indicator has been observed and (partially) explained by Rousseau (2007) and Vanclay (2007).

Yet, it is well-known that the h-index has several disadvantages. Some of these it shares with any other citation-based indicator.

- Like most pure citation measures it is field-dependent, and may be influenced by self-citations.
- There is a problem finding reference standards.
- There exist many more versatile indicators for research evaluation.
- The number of co-authors may influence the number of citations received.
- It is rather difficult to collect all data necessary for the determination of the h-index. Often a scientist's complete publication list is necessary in order to discriminate between scientists with the same name and initial. We refer to this problem as the precision problem.

Comments: Iglesias and Pecharromán (2007) proposed a correction factor (based on the Essential Science Indicators) in order to compare h-index calculations in different fields. We would like to remark though that we do not favour the idea of cross-field comparisons. The relation between the h-index and self-citations has been studied by Schreiber (2007). Not surprisingly, the influence of self-citations on the h-index goes from negligible to considerable. The pure h-index (Wan et al., 2007) takes the number of co-authors and possibly their contribution to the article into account, as does Schreiber's h_m (Schreiber, 2008) and the adapted pure h-index (Chai et al., 2008). In the first approach *citations* are counted fractionally, while in the second approach *articles* are counted fractionally. A mathematical approach to fractional counting of publications as well as of citations has been written by Egghe (2008a). The precision problem will play an important role in this review when it comes to decide between different proposals aiming at removing some of the disadvantages of the h-index in its original form.

Some of the disadvantages are more specifically related to the h-index itself.

- The h-index, in its original setting, puts newcomers at a disadvantage since both publication output and observed citation rates will be relatively low. In other words, it is based on long-term observations.
- The index allows scientists to rest on their laurels since the number of citations received may increase even if no new papers are published.
- The h-index is only useful for comparing the better scientists in a field. It does not discriminate among average scientists.
- The h-index lacks sensitivity to performance changes: it can never decrease and is only weakly sensitive to the number of citations received.

This last point is the reason for the introduction of most other h-type indices.

Comments: The fact that the h-index puts newcomers at a disadvantage is only true for the original context of the h-index as a life-time achievement index. In this context this not a disadvantage but just a part of the aim of the index. However, if the h-index is not meant as a life-time achievement index, but, for instance, is used to evaluate the contemporary quality of a scientist, then the h-index must be defined in a slightly different way. This can be done, for instance, by restricting publications to the latest Y years. A solution for the fact that the h-index never decreases and that scientists can 'rest on their laurels' is provided by Jin's AR-index (Jin, 2007). The h-index can be refined, yielding real or rational values, hence discriminating among average scientists. This is discussed in the next section.

3 Refinements of the h-index: the real-valued and the rational h-index

3.1 The real-valued h-index, h_r (Rousseau, 2006b)

Let P(r) be the production of the r^{th} source in a ranked list of articles according to citations and let P(x) be its piecewise linear interpolation, then h_r is the intersection of the function P(x) and the line y = x. This approach seems most useful when the number of citations is also a real or rational number, which is the case when citations are counted fractionally. Clearly: $h \leq h_r < h+1$.

3.2 The rational h-index, h_{rat} (Ruane & Tol, 2008)

This h-index, denoted as h_{rat}, also satisfies the inequality $h \leq h_{rat} < h+1$. It is defined as (h+1) minus the relative number of scores necessary for obtaining a value h+1.

Let n be the number of citations necessary for obtaining an h-index one higher. This number n is divided by the highest possible n, namely 2h+1. Indeed, the lowest possible situation leading to a h-index of h consists of h articles with h citations, followed by an article without any citation. In order to get an h-index equal to h+1 one needs one more score for each of the first h sources, h scores in total, and h+1 scores for the last one: a total of 2h+1.

Example a. The ranking 3 – 3 – 3 – 0 (articles ranked according to the number of citations) leads to h = 3. Its $h_{rat} = 4 - 7/7 = 3$.

Example b. The ranking 4 – 4- 4 – 3 has h = 3; its $h_{rat} = 4 - 1/7 = 27/7$, or almost 4.

Example c. The ranking x – y- z – 3 with $x \geq y \geq z \geq 4$ has the same h_{rat} as example 4, showing that, as the original h, also h_{rat} is not sensible to the exact number of citations (and neither is h_r).

The rational h-index seems an intelligent and useful extension to the original h-index. It helps to overcome the problem of ranking scientists with the same (classical) h-index. We recommend its use.

4 Generalizations of the Hirsch index

Because the h-index is only weakly sensitive to the actual number of citations received Egghe proposed another index, referred to as the g-index (Egghe, 2006). The g-index is calculated as follows: one draws the same list as for the h-index, but now the g-index is defined as the highest rank such that the cumulative sum of the number of citations received is larger than or equal to the square of this rank. Clearly $h \leq g$. The g-index has the disadvantage that it needs more data than the h-index and hence aggravates the precision problem.

This disadvantage is overcome by another h-type index, called the R-index (Jin et al. 2007). As a mathematical formula the R-index is defined as:

$$R = \sqrt{\sum_{j=1}^{h} cit_j} \qquad (1)$$

In formula (1) the numbers of citations (cit_j) are ranked in decreasing order. Note, that, as long as the h-core contains exactly h elements, the R-index is unambiguously defined (one does not need a specific definition for the h-core). It is also clear that $h \leq R$. The R-index, moreover, uses the same data as the h-index so that the precision problem is exactly the same as for the original h-index, and is not increased as was the case for the g-index. For this reason we consider this the best index among those that take citations into account.

Other h-type indices include the pure h-index (Wan et al., 2007), Schreiber's h_m (Schreiber, 2008) mentioned above and the tapered h-index (Anderson et al., 2008).

Jin Bihui introduced the AR-index (or its square the AR^2-index) in order to incorporate the age of published articles (Jin, 2007). This is a good idea as, clearly, for research evaluation purposes work performed twenty years ago is of less importance than work performed four years ago. The AR-index is defined as:

$$AR = \sqrt{\sum_{p \in H} \frac{cit_p}{a_p}} \quad (2)$$

In this formula, H denotes the h-core and p denotes a publication; cit_p denotes the number of citations received by publication p; a_p denotes the age of article p. Recent investigations (Rousseau & Jin, 2008) seem to suggest that it is somewhat better to remove the square root and use AR^2 instead.

Two scientists can have the same h-index, and even the same R-index but one's career can be on the rise (citation-wise), while the other one's is stagnating. A so-called dynamic index taking this aspect into account is proposed by Rousseau and Ye (2008).

5 Use of the h-index

The h-index can not only be used for lifetime achievements (as Hirsch did), but also in the context of many – but not all - other source-item relationships. Consequently, the Hirsch index has been calculated for other periods than a scientist's lifetime, journal citations (Braun et al., 2006; Rousseau, 2006a), topics (Banks, 2006; STIMULATE, 2007) and library loans per category (Liu & Rousseau, 2007),

Mugnaini et al. (2008) use the h-index to compare scientists of the Brazilian Academy of Sciences with their counterparts in the USA. It is not surprising that the Americans stand out. Yet, among Brazilian scientists physicists and mathematicians perform relatively better.

In *Does the h-index have predictive power?* Hirsch (2007) shows that the h-index is better than total citations, citations per paper and total number of publications in predicting future scientific achievements. In this article he also disproves claims by a group of Denmark-based researchers (Lehmann et al., 2005) who argue that the mean number of citations per year is a superior indicator of scientific quality.

6 Prathap's approach to research evaluation and successive h-indices

In a brief letter published in *Current Science*, Prathap (2006) proposed two levels for using the *h*-index in institutional evaluations, taking into account a first order *h*-index (h_1) and a second order *h*-index (h_2), where the institute's first order *h*-index is equal to h_1 if the institution (this is the group of all its researchers) has published h_1 papers, each of which has at least h_1 citations; and its second order *h*-index is h_2 if the institution has h_2 researchers, each having an individual *h*-index which is at least equal to h_2. The calculation of Prathap's h_2 is a special case of Schubert's idea of successive *h*-indices (Schubert, 2007). Prathap's main idea is to compare h_1 and h_2, using them as complementary indices. Indeed: h_1 can be high because the institute has many researchers that are highly cited, or because the institute has just a few scientists with a very high h-index. The difference between the two cases is made clear by h_2. Arencibia-Jorge and Rousseau (2008) present an example of the application of Prathap's method. They study the twelve most productive Cuban institutions in the field of human brain research.

In a somewhat similar vain Egghe and Rao (2008) consider several h-indices for a group of authors, e.g. those working at the same institute. They study and model the h-index of all publications, denoted as h_P, reflecting the *authors – number of publications* relation; the h-index of all citations received, denoted as h_C, reflecting the *authors-number of citations* relation; and Prathap's h_1 (denoted as h_G by Egghe & Rao) and h_2 indices. These indices reflect the *articles-citations* relation and the *authors versus h-indices* relation. They show that in the power law model $h_2 < h_P < h_C$ and $h_2 < h_1$. We observe that they could have added an h-index reflecting the *authors – number of collaborators* relation.

7 The influence of the database used for calculation an h-type index

Judit Bar-Ilan (2008) compared h-indices of highly cited Israeli researchers based on citation counts retrieved from WoS, Scopus and Google Scholar. Often, results obtained through Google

Scholar differed considerably from results obtained from the other sources. Moreover, the precision problem for Google Scholar becomes very hard as illustrated by Meho and Yang (2007).

8 Time series of journal h-indices: definitions

As it is utterly impossible that one number such as the h-index fully characterizes a scientist or a journal one is often not so much interested in the h-index itself, or any other h-type index, but more how this index changes over time. This leads to the idea of series of h-indices. A study of the change of this number over time comes one step closer to a complete description, and still has the advantage of being a summary statistic. The first to study series of h-indices was Liang Liming (2006). Rousseau (2006a) considered a time series of h-indices for the journal JASIS and suggested that for the calculation of journal h-indices a normalization with respect to the number of published articles would be appropriate. Burrell (2007a), based on a linear relation between the h-index and age (career age in the case of life-time achievements), proposed the h-rate as an alternative indicator.

Many different types of series of h-indices are possible. Liu & Rousseau (2008) give ten examples and propose a notation for unambiguously describing the type of h-series actually used or studied. They also use a clear visualization for time series of citation data based on a publication-citation matrix (Ingwersen et al., 2001). Examples of series of h-indices are given in Liu et al. (2008).

9 Mathematical models

Hirsch (2005) proposed a baseline model in which the number of publications is the same each year, and each published article receives the same number of citations each year. Under this model h increases linearly with age.

Egghe (2007a,b), using a power law model for the citation function, this is the number of articles receiving n citations, found that if the citation function at time t = ∞ (the asymptotic power law) is given then the corresponding power law at a finite time t can be found. From this knowledge the h (and g) -index at time t can be derived. He showed that these time-dependent functions are concavely increasing with a horizontal asymptote. During recent years he wrote many articles applying his power law model to the h-index and h-type indices. Also Burrell (2007b) devised models describing the h-index and its generalizations. Glänzel (2008a) studies statistical properties of the h-index in the context of extreme-value statistics. Merging or concatenation of publication lists and the effect on the resulting h-index has been studied by Egghe (2008b) and Glänzel (2008b).

10 Other developments: statistics and axiomatics

The h-index can be considered a stochastic variable defined using a sample (a scientist's publications). Each publication is assumed to have the same citation distribution function, which is either discrete or continuous. In Egghe's publications this citation distribution function is always a power law, i.e. a Pareto distribution. Clearly the h-index depends on the sample size. Interesting statistical articles are Burrell's (2007b), Glänzel's (2008a,b) and Beirlant & Einmahl's (2007).

A very interesting new line of investigation was introduced by Gerhard Woeginger (2008a). He takes an axiomatic approach to the h-index and h-type indices. Representing a researcher with $n \geq 0$ publications by a publication-citation array $x = (x_1, \ldots, x_n)$ with non-negative integer-valued components $x_1 \geq x_2 \geq \ldots x_n$, where x_k denotes the number of citations received by the k-th publication (during the period under investigation), he defines a scientific impact index as a function f from the set X of all publication-citation arrays into the set $\mathbb{N}$ of non-negative integers satisfying the following three axioms. Note that also the empty array belongs to X.

- If x is the empty array then f(x) = 0.

- If $x = (x_1, \ldots, x_n)$ and $y = (x_1, \ldots, x_n, 0)$ then $f(x) = f(y)$.
- Monotonicity: if $x = (x_1, \ldots, x_n)$, $y = (y_1, \ldots, y_m)$ and $x_k \leq y_k$ for $1 \leq k \leq n = \min(n,m)$, then $f(x) \leq f(y)$.

Woeginger (2008a) succeeds in finding additional axioms characterizing the h-index, the g-index (Woeginger, 2008b) and some other indices.

11 A classification

Some authors have tried to refine h-index calculations; others use combinations of the h-index with a secondary index, while a third group wants to replace the h-index by a 'better' one. This is schematically shown in Figure 1.

12 Conclusion

The study of the h-index, its generalizations and its applications is a fruitful field of investigation. It is our conviction that, besides the h-index itself, the statistical and the axiomatic lines of investigation show the most promise for having a long-lasting influence on the field of informetrics. Woeginger's articles are certainly the most exciting development of the latest months.

Acknowledgement

Research of R. Rousseau was supported by the National Natural Science Foundation of China through grant no. 70673019.

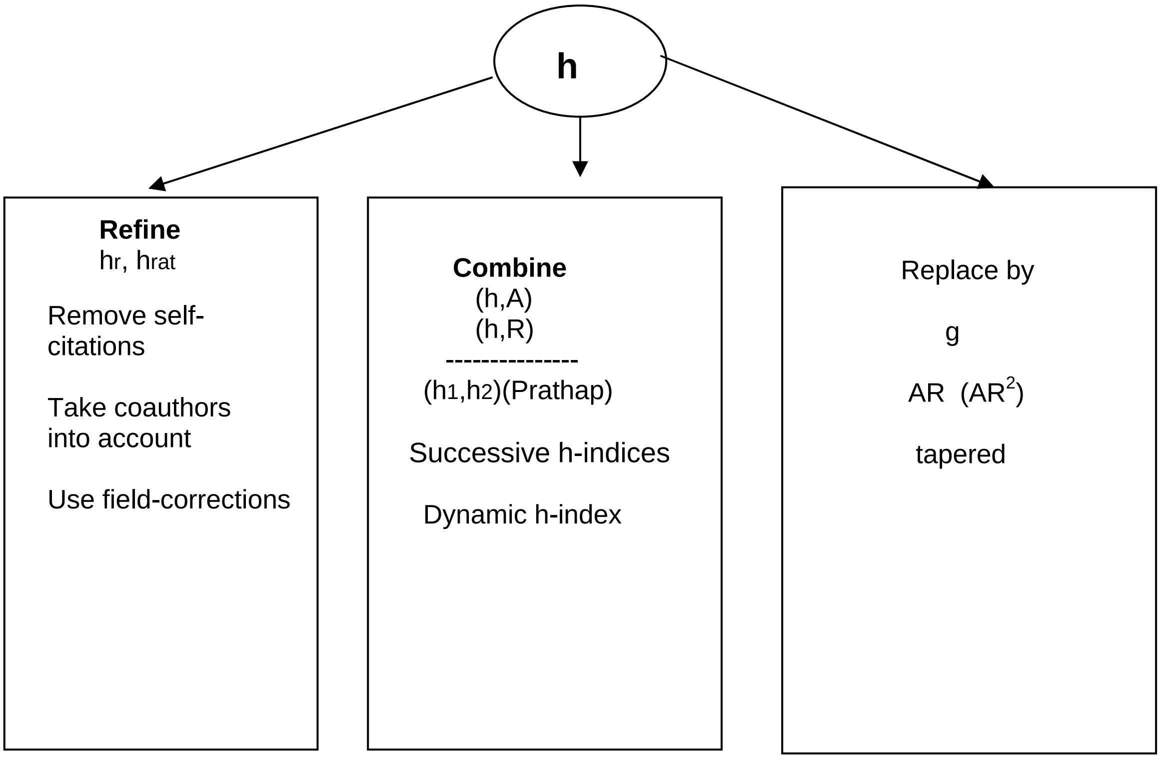

Figure 1

References

Anderson, T.T., R.K.S. Hankin and P.D. Killworth (2008). Beyond the Durfee square: enhancing the h-index to score total publication output. *Scientometrics* (to appear).

Arencibia-Jorge, R. and R. Rousseau (2008). Influence of individual researchers' visibility on institutional impact: an example of Prathap's approach to successive h-indices. *Scientometrics* (to appear).

Ball, P. (2007). Achievement index climbs the ranks. *Nature*, 448, p. 737.

Banks, M.G. (2006). An extension of the Hirsch index: indexing scientific topics and compounds. *Scientometrics*, 69, 161-168.

Bar-Ilan, J. (2008). Which h-index? A comparison of WoS, Scopus and Googlr Scholar. *Scientometrics.* 74, 257-271.

Beirlant J. and J.H.J. Einmahl (2007). Asymptotics for the Hirsch index. CentER Discussion paper (2007-86). Tilburg University.

Bornmann, L. and H.-D. Daniel (2007). What do we know about the h index? *Journal of the American Society for Information Science and Technology,* 58, 1381-1385.

Bormann. L., R. Mutz and H.-D. Daniel (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, 830-837.

Braun, T., W. Glänzel and A. Schubert (2006). A Hirsch-type index for journals. *Scientometrics*, 69, 169-173.

Burrell, Q.L. (2007a), Hirsch index or Hirsch rate? Some thoughts arising from Liang's data. *Scientometrics*, 73: 19-28.

Burrell, Q.L. (2007b). Hirsch's h-index and Egghe's g-index. In: *Proceedings of ISSI 2007* (Daniel Torres-Salinas & Henk F. Moed, eds.). Madrid: CINDOC-CSIC, pp. 162-169.

Chai, JC., PH Hua, R. Rousseau and JK Wan (2008). The adapted pure h-index. *These proceedings*.

Egghe, L. (2006). An improvement of the h-index: the g-index. *ISSI Newsletter*, 2(1), 8-9.

Egghe, L. (2007a). Dynamic h-index: the h-index in function of time. *Journal of the American Society for Information Science and Technology*, 58, 452-454.

Egghe, L. (2007b). Item-time-dependent Lotkaian informetrics and applications to the calculation of the time-dependent h-index and g-index. *Mathematical and Computer Modelling*, 45, 864-872.

Egghe, L. (2008a). Mathematical theory of the h- and g-index in case of fractional counting of authorship. *Journal of the American Society for Information Science and Technology,*

Egghe, L. (2008b). The influence of merging on h-type indices. *Journal of Informetrics* (to appear).

Egghe, L. and I.K.R. Rao (2008).Study of different h-indices for groups of authors. *Journal of the American Society for Information Science and Technology*, 59, 1276-1281.

Glänzel, W. (2006). On the opportunities and limitations of the h-index. *Science Focus*. 1(1), 10-11. (In Chinese). An English translation can be found at E-LIS ID-code 9535.

Glänzel, W. (2008a). On some new bibliometric applications of statistics related to the h-index. *Scientometrics*, 76(3) (to appear).

Glänzel, W. (2008b). H-index concatenation. Preprint.

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.

Hirsch, J. E. (2007). Does the h-index have predictive power? *Proceedings of the National Academy of Sciences of the United States of America,* vol. 104 (49), 19193-19198.

Iglesias, J.E. and C. Pecharromán (2007). Scaling the h-index for different scientific ISI fields. *Scientometrics*, 73, 303-320.

Ingwersen, P., B. Larsen, R. Rousseau and J.M. Russell (2001). The publication-citation matrix and its derived quantities. *Chinese Science Bulletin*, 46: 524-528.

Jin, BH. (2007). The AR-index: complementing the h-index. *ISSI Newsletter,* volume 3, number 1, p.6.

Jin, BH., LM. Liang, R. Rousseau and L. Egghe (2007). The R- and AR-indices: complementing the h-index. *Chinese Science Bulletin*, 52, 855-863.

Lehmann, S., A.D. Jackson and B.E. Lautrup (2005). Measures and mismeasures of scientific quality. arXiv:physics/0512238.

Liang, LM. (2006). H-index sequence and h-index matrix: constructions and applications. *Scientometrics*, 69, 153-159.

Liu, YX., I.K.R. Rao and R. Rousseau (2008). Empirical series of journal h-indices: the JCR category *Horticulture* as a case study. *Scientometrics* (to appear).

Liu, YX. and R. Rousseau (2007). Hirsch-type indices and library management: the case of Tongji University Library. In: *Proceedings of ISSI 2007* (Daniel Torres-Salinas & Henk F. Moed, eds.). Madrid: CINDOC-CSIC, pp. 514-522

Liu, YX and R. Rousseau (2008). Definitions of time series in citation analysis with special attention to the h-index. *Journal of Informetrics* (to appear).

Meho, L.I. and K. Yang (2007). Impact of data sources on citation counts and rankings of LIS faculty: Web of Science versus Scopus and Google Scholar. *Journal of the American Society for Information Science and Technology*, 58(13), 2105-2125.

Mugnaini, R., A.L. Packer and R. Meneghini (2008). Comparison of scientists of the Brazilian Academy of Sciences and of the National Academy of Sciences of the USA on the basis of the h-index. *Brazilian Journal of Medical and Biological Research*, 41, 258-262.

Prathap, G. (2006) Hirsch-type indices for ranking institutions' scientific research output. *Current Science*, 91, p. 1439.

Rousseau, R. (2006a). A case study: evolution of JASIS' h-index. *Science Focus*, 1(1): 16-17 (in Chinese). English version available at E-LIS: ID 5430.

Rousseau, R. (2006b). Simple models and the corresponding h- and g-index. E-LIS: ID 6153.

Rousseau, R. (2007). The influence of missing publications on the Hirsch index. *Journal of Informetrics*, 1, 2-7.

Rousseau, R. and BH. Jin (2008). The age-dependent h-type AR²-index: basic properties and a case study. Preprint.

Rousseau, R. and F.Y. Ye (2008). A proposal for a dynamic h-type index. *Journal of the American Society for Information Science and Technology*,59(10) (to appear)

Ruane, F.P. and R.S.J. Tol (2008). Rational (successive) H-Indices: An application to economics in the Republic of Ireland. *Scientometrics*, 75, 395-405.

Schreiber, M. (2007). Self-citation corrections for the Hirsch index. *EPL*, 78, 30002.doi:10.1209/0295-5075/78/30002

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.

Schubert, A. (2007). Successive h-indices. *Scientometrics,* 70, 201-2005.

The STIMULATE-6 Group (2007). The Hirsch index applied to topics of interest to developing countries. *First Monday*, 12(2). http://www.firstmonday.org/issues/issue12_2/stimulate/

Vanclay, J.K. (2007), On the robustness of the h-index. *Journal of the American Society for Information Science and Technology,* 58, 1547-1550.

Wan, JK, PH Hua R. Rousseau (2007). The pure h-index : calculating an author's h-index by taking co-authors into account. *COLLNET*

Journal of Scientometrics and Information Management, vol. 1 (2), 1- 5.

Woeginger, G.J. (2008a). An axiomatic characterization of the Hirsch-index. *Mathematical Social Sciences* (to appear).

Woeginger, G.J. (2008b). An axiomatic analysis of Egghe's g-index. *Journal of Informetrics* (to appear).

H. Kretschmer & F. Havemann (Eds.): Proceedings of WIS 2008, Berlin
Fourth International Conference on Webometrics, Informetrics and Scientometrics & Ninth COLLNET Meeting
Humboldt-Universität zu Berlin, Institute for Library and Information Science (IBI)