

## **The plausibility of computing the h-index of scholarly productivity and impact using reference enhanced databases**

The h-index, developed by Jorge E. Hirsch to quantify the scientific output of researchers, has immediately received well-deserved attention in academia. The theoretical part of his idea was widely embraced, and even enhanced by several researchers. Many of them also recommended derivative metrics based on Hirsch's idea to compensate for potential distortion factors, such as high self-citation rates. The practical aspects of determining the h-index also need scrutiny because some content and software characteristics of reference enhanced databases can strongly influence the h-index values arrived at. This paper provides a general overview of these issues, to be followed by a series of papers focusing on the analysis of pros and cons of the three largest, cited reference enhanced, multidisciplinary databases: Google Scholar, Scopus and Web of Science for determining the h-index.

Bibliometric, informetric and scientometric measures are being used more and more extensively for evaluating scientists, researchers, universities, research institutions and their departments at the national and international levels, among others in decisions for awarding tenure, grants, and allocating research funds. Such metrics form an integral part of the research assessment and evaluation in the UK, Australia, New Zealand, and several other countries. These have been almost exclusively based on the citation indexes of the Institute for Scientific Information (now Thomson Scientific).

The recently introduced new bibliometric measure, the h-index, (Hirsch, 2005) was developed to characterize through a single number, both the productivity (number of publications), and the impact or influence (citedness of publications) of scholars. It immediately generated immense interest among researchers and academic administrators. This was well-deserved because of the convincing and transparent theoretical foundation, and the practical simplicity of creating h-index lists from simple database searches not only for individuals, but also for research groups, institutions and journals.

Beyond the simplicity and convenience, however, considerations must be given also to the prevailing software and content limitations of the databases, and to the peculiarities of the h-index generating utility programs that are used in the process. The lack of necessary knowledge about the content and software, and the search skills of the persons who determine the h-index may aggravate the situation.

The h-index is generated from a set of results produced through queries, performed by searchers of very different search skills. They use a variety of databases that can report the citedness of the works of an author represented by master records in the result lists. This process is different from the calculation of the controversial but widely used journal impact factor, which is also a single-number bibliometric measure based on citation analysis, prepared centrally from a partially normalized dataset by highly qualified specialists and reported annually in the Journal Citation Reports of Thomson Scientific. As long as the software can report at least the citation counts for each items, and preferably can sort the result by citation count, lists of h-indexes may be created using several different databases and database editions (in case of Web of Science, and some Ebsco databases), choosing different options for generating the h-index (in case of Scopus), or different third-party programs (in case of Google Scholar).

While there are several databases which may be ideal not only for discovering and identifying relevant documents for a research topic or by an author, journal or institute, but also searching the full text and delivering the source documents, they may be entirely inappropriate for computing the h-index. This is obvious when a database does not offer citation counts for its bibliographic master records as is the case with the excellent Psychology and Behavioral Sciences databases of EBSCO, which has full text records for 85% of its nearly 700,000 item database, but does not (yet) provide information about the citedness of items, let alone offers citation searching.

In other cases, however, when the result list does include such information and the software offers citation searching options, it may not be obvious that the database is still inappropriate for h-index purposes, as is the case with many of the otherwise worthy databases of Ebsco where bibliographic records are enhanced with cited references only very selectively, and only for the past few years.

This introductory paper explains the essential practical requirements for creating h-index lists, and the considerations in interpreting the results, as they may be highly distorted due to a) shortcomings in the breadth (retrospectivity, depth and width), and consistency of coverage in the databases, and b) potential deficiencies in the citation matching algorithm and in the handling of the quasi matches and non-matches by the software.

A series of subsequent papers will present the specific pros and cons of the three largest, multidisciplinary databases enhanced by or based on cited references: Google Scholar, Scopus, and Web of Science for determining and interpreting the h-index, and the reasons for considerable differences for the same researchers. Judit Bar-Ilan (2007) sums up the most succinctly the problem in the title (Which h-index?) of her recent paper about the h-indexes produced by Web of Science, Scopus and Google Scholar for a group of prominent Israeli scientists.

### **The concept and caveats of the h-index**

Jorge E. Hirsch (2005) summarized the h-index as “the number of papers with citation number  $\geq h$ , as a useful index to characterize the scientific output of a researcher”. For example, the researcher J. E. Hirsch has a h-index of 50, if he has at least 50 papers which were cited at least 50 times.

Hirsch demonstrates his new measure showing the h-index of 21 most prominent physicists, from E. Witten ( $h=112$ ) to S.W. Hawking ( $h=62$ ). (The latter is in tie with M.S. Dresselhaus who was reported in Hirsch’s paper by a significantly lower  $h=62$  index than she would have deserved according to my calculations. As of early February, 2008 Hawking’s h-index is still 62, while Dresselhaus has  $h=76$ , and it is not caused by a sharp surge in the citations received by her works in the past two years).

Hirsch had a novel idea and presented it simply and convincingly, in PNAS (*Proceedings of the National Academy of Sciences of the USA*), one of the most influential academic periodicals. (Quite tellingly, this short paper, which probably has the most succinct abstract, and must be one of the few papers whose author dares to use first person singular in PNAS, was cited over 100 times within two years).

There are a number of papers that suggest refinements, and enhancements of the original idea (Egghe, 2006; Jin et al, 2007; Vanclay, 2006), or extend its application beyond individuals (Banks, 2006, Braun et al., 2006; Prathap, 2005), and only a very few that had negative reactions to it (Berger, 2007; Purvis, 2006). There are numerous papers about applying and testing the idea of Hirsch in various disciplines (Bar-Ilan, 2007; Cronin and Meho, 2006; Harzing and van der Wal, 2008; Meho and Yang, 2007; Oppenheim, 2007, Schreiber, 2007a, Saad, 2006).

Hirsch mentions some of the caveats that must be applied in interpreting the h-index, and he duly notes that it should be used together with other assessment criteria. Among his

caveats he warns that “there will be differences in typical h values in different fields, determined in part by the average number of references in a paper in the field, the average number of papers produced by each scientists in the field, and the size (the number of scientists) of the field”.

Hirsch aptly demonstrates this by showing the h-index of 10 of the most prominent researchers in the biosciences from S.H. Snyder (h=191) to A. Ullrich (h=120). The lowest h-index in the bioscientist group is higher than the highest one among the physicists, and almost twice as high as the lowest one in physics. The median h among the top 10 physicists and bioscientists is 75, and 147, respectively. The difference would be larger if the calculation of the h-index for the group of bioscientists would not have been restricted to 1983-2005, but calculated from 1955, as it was done for the physicists.

Hirsch elaborates on the issue, warning that “scientists working in non-mainstream areas will not achieve the same very high h values as the top echelon of those working in highly topical areas”. In my experience there are significant differences even within a discipline or sub-discipline, such as cataloging digital resources versus cataloging Eastern European incunabula. Finally, he brings up the issue of self-citation which was the topic of several of the articles triggered by the introduction of the h-index (Schreiber, 2007b, Vinkler, 2007). Hirsch admits that self-citations can increase the individual’s h-index, but he argues that self-citations would have far less impact on the h-index than on the measure of total number of citations received. He concludes saying that “I have proposed an easily computable index, h, which gives an estimate of the importance, significance, and broad impact of a scientist’s **cumulative** (emphasis mine) research contributions”.

There is also a warning – although only in a tiny footnote of 8-point size, that “[o]f course, the database used must be complete enough to cover the full period spanned by the individual’s publications”. Completeness is just one of my main concerns when putting the theory into practice, and not only for the reason that completeness of the database is a vary vague issue for most users, but also because it is not easy to realize the other content and software limitations of the database(s) that are used to determine the h-index. These non-obvious limitations may grossly distort the results.

### Different h-index from different databases

It is just natural to lead in this discussion with the question: what is the h-index of J.E. Hirsch? The answer depends on which database(s) one uses, and how. Table 1 shows the huge differences in the h-indexes computed from eight databases, ranging from the impressively high h-index in WoS for a researcher with 30 years of publishing activity to the absurdly low h-index in the very large multidisciplinary databases of Ebsco.

Database	h-index	Master records	Total citations	Avg citations
WoS (Web of Science)	51	179	9 288	51.888
ADS (Astrophysics Data System)	41	181	6 633	36.646
Google Scholar	38	219	6 191	28.269
PROLA (Physical Review Online Archive)	36	127	4 912	38.677
Scopus	33	168	3 938	23.440
arXiv	14	49	719	14.673
Ebsco MegaFILE	1	10	4	0.400
Academic Search Premier	0	16	0	0

Table 1. The h-indexes for physicist Jorge E Hirsch from 8 databases

In Web of Science (WoS), Hirsch's h-index is 51, based on 9,288 references to 179 publications by "our" J. E. Hirsch, in periodicals covered by WoS. It is to be noted that 6 records for papers by J. E. Hirsch published before 1980, are not included in the 1980-2008 edition of WoS used for this test, but they don't change the h-index, even if it had been calculated from the WoS editions starting their coverage from 1945 or 1955. Hirsch started publishing in the mid-1970s, and none of his first 6 papers have been cited more than 51 times. The term "edition of WoS" is just my practical shorthand, not an official term, and it depends on the libraries' decision how many years of coverage the library chose when it licensed the WoS database.

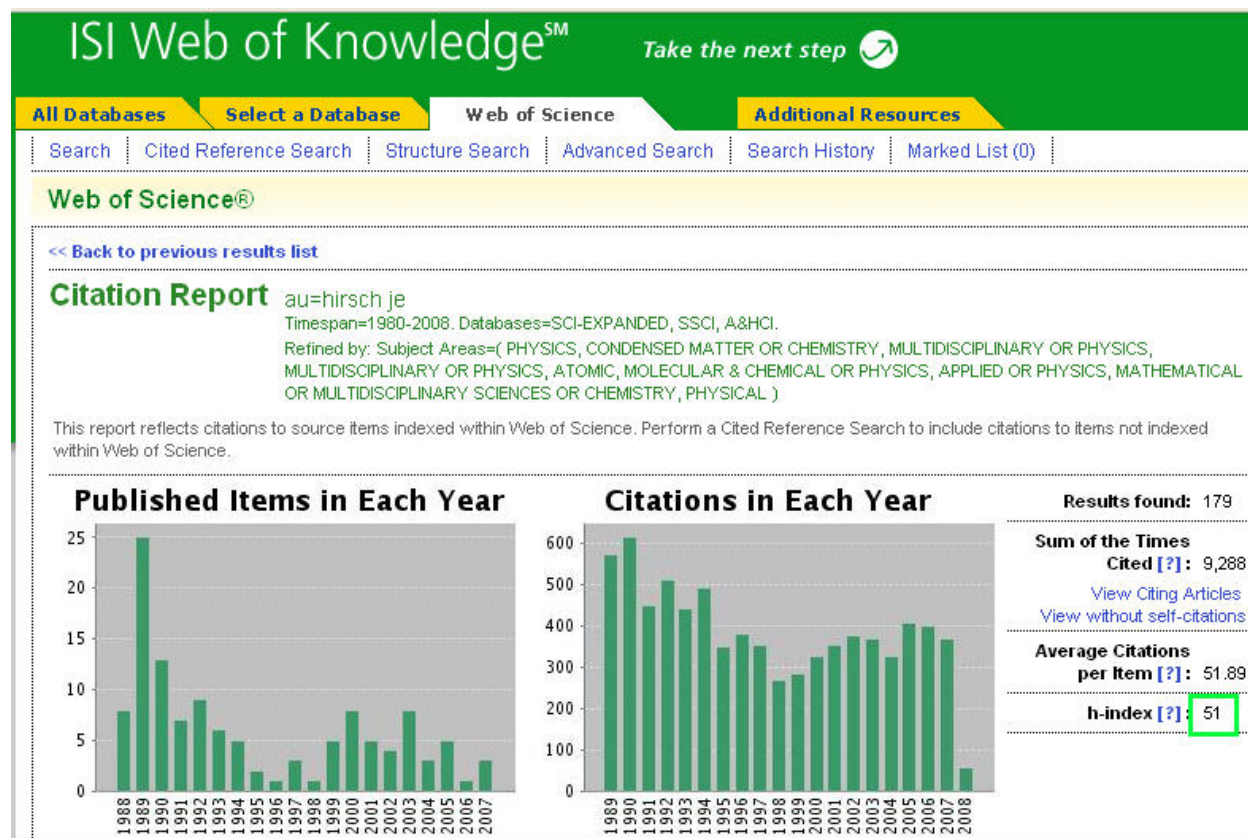


Figure 1. The Citation Report feature of WoS shows -among other informative bibliometric measures- the h-index

In Scopus the h-index of J.E. Hirsch is 33, based on 3,938 references to 168 publications by "our" J. E. Hirsch in sources covered by Scopus. In WoS, Scopus and Google Scholar the searches were limited to physics and other disciplinary areas for filtering the set for Hirsch's source papers produced by the name and initials combination. This eliminates records for papers by authors in other disciplines with the same name and initials combination, the ones who are not "our" J. E. Hirsch. In the astronomy and physics databases this is implied. In cases where two researchers have the same family name and middle initial(s), and work in the same subject field, this problem must be handled in a tedious way. That's the reason that in many h-index lists 10-15% of the originally considered potential researchers' h-indexes are eliminated or not calculated. This was the case also in Hirsch's tests.

Scopus: 168   More... (0)   Web (49)   Patents (0)

Your query: AUTHOR-NAME(hirsch,j e) AND (LIMIT-TO(SUBJAREA, "PHYS") OR LIMIT-TO(SUBJAREA, "MATE") OR LIMIT-TO(SUBJAREA, "MULT") OR LIMIT-TO(SUBJAREA, "MULT"))   [Search History](#)

[Edit](#)   [Save](#)   [Save as Alert](#)   [RSS](#)

**Refine Results** [Open](#)

**Results: 168**   Search within results  [Go](#)

[Output](#)   [Citation tracker](#)   [Add to list](#)

[References](#)   [Cited by](#)   Select:  All  Page 1 to 50 [Next](#)

Document (sort by relevance)	Author(s)	Date	Source Title	Cited By
1. <input type="checkbox"/> <b>Spin Hall effect</b> <a href="#">Abstract + Refs</a> <a href="#">View at Publisher</a> <a href="#">Show Abstract</a>	<a href="#">Hirsch, J.E.</a>	1999	<i>Physical Review Letters</i> 83 (9), pp. 1834-1837	<a href="#">289</a>
32. <input type="checkbox"/> <b>Pairing interaction in two-dimensional CuO<sub>2</sub></b> <a href="#">Abstract + Refs</a> <a href="#">View at Publisher</a> <a href="#">Show Abstract</a>	<a href="#">Hirsch, J.E., Tang, S., Loh Jr., E., Scalapino, D.J.</a>	1988	<i>Physical Review Letters</i> 60 (16), pp. 1668-1671	<a href="#">33</a>
33. <input type="checkbox"/> <b>Efficient Monte Carlo procedure for systems with fermions</b> <a href="#">Abstract + Refs</a> <a href="#">View at Publisher</a> <a href="#">Show Abstract</a>	<a href="#">Hirsch, J.E., Scalapino, D.J., Sugar, R.L., Blankenbecler, R.</a>	1981	<i>Physical Review Letters</i> 47 (22), pp. 1628-1631	<a href="#">33</a>

Figure 2. Scrolling -down in the result list sorted by citedness –is one of several options for determining the h-index in Scopus

In Google Scholar the h-index is 38 (based on 6,191 references to 219 records for papers attributed to J. E. Hirsch by Google Scholar). It is to be noted that in Google Scholar the master records and the orphan reference records (to be discussed later) are merged in the result lists. The hit counts and the citation counts of Google Scholar keep changing dramatically. If they were increasing it could be chalked up to adding new records, but often these counts decrease because of deleting records from the database (see the example of disappearing records after I demonstrated an extreme mismatch-making at <http://www2.hawaii.edu/~jacso/extra/>)

In the Astrophysics Data System (ADS) the h-index of J. E. Hirsch is 41 based on 6,642 references (6,633 at my original search, but the increase did not change the h-index value) to 181 master records for papers of Hirsch available in this database. The Physics Review Online Archive (PROLA) of the American Physical Society, which recently introduced the feature of sorting the search results by the papers' citedness, yields a h-index of 36 based on 4,912 references to 127 records for papers that Hirsch published in journals of the APS. There are two reasons for this surprisingly high but realistic h-index from the digital archive of a single publisher.

One is that Hirsch wrote more than 90% of his papers for one of the publications of APS. The other reason is that the number of citing references in PROLA is not limited to sources published by APS, but includes citations from articles in journals of other publishers who participate in the CrossRef project which was created to facilitate inter-

publisher linking to cited and citing references. This is an important development because (as of early January, 2008) more than 2,500 publishers participate in CrossRef with nearly 20,000 journals and 31 million articles. Some other publishers, such as Springer, also started to use CrossRef to report the citedness of papers not only from their own journals, but also from journals of other publishers which are CrossRef members. Unfortunately, CrossRef is not searchable directly by author name for end-users, neither does it report the citedness of the papers. The same is true for the new interface of the digital archive of the Institute of Physics, which seems to be a step backward. The American Institute of Physics' Scitation system covers several publishers' journals, and it does offer searching by author name (the taken for granted option) but does not report the citedness of articles. From the perspective of the h-index, this is a limitation of several other online information services, such as Scitopia, and WorldWide Science, which would have a broader base of coverage as they combine records for papers published in journals of several large publishers. Currently, none of these services and databases can be used for calculating the h-index.

[SAO/NASA Astrophysics Data System \(ADS\)](#)

Query Results from the ADS Database

[Go to bottom of page](#)

Selected and retrieved 181 abstracts. Total citations: 6633

#	Bibcode Authors	Cites Title	Date	<a href="#">List of Links</a> <a href="#">Access Control Help</a>		
1	<input type="checkbox"/> <a href="#">1985PhRvB..31.4403H</a> Hirsch, J. E.	391.000 Two-dimensional Hubbard model: Numerical simulation study	04/1985	<a href="#">A</a> <a href="#">E</a>		
2	<input type="checkbox"/> <a href="#">1985PhRvL..54.1317H</a> Hirsch, J. E.	319.000 Attractive interaction and pairing in fermion systems with strong on-site repulsion	03/1985	<a href="#">A</a> <a href="#">E</a>	<a href="#">R</a> <a href="#">C</a>	
.....						
39	<input type="checkbox"/> <a href="#">1991PhRvB..43.424H</a> Hirsch, J. E.; Marsiglio, F.	43.000 Hole superconductivity in oxides: A two-band model	01/1991	<a href="#">A</a> <a href="#">E</a>	<a href="#">R</a> <a href="#">C</a>	
40	<input type="checkbox"/> <a href="#">1985PhRvB..31.6022H</a> Hirsch, J. E.	42.000 Phase diagram of the one-dimensional molecular-crystal model with Coulomb interactions: Half-filled-band sector	05/1985	<a href="#">A</a> <a href="#">E</a>	<a href="#">R</a> <a href="#">C</a>	
41	<input type="checkbox"/> <a href="#">1989PhRvB..40.5000T</a> Tang, Sanyee; Lazzouni, M. E.; Hirsch, J. E.	41.000 Sublattice-symmetric spin-wave theory for the Heisenberg antiferromagnet	09/1989	<a href="#">A</a> <a href="#">E</a>	<a href="#">R</a> <a href="#">C</a>	<a href="#">U</a>

Figure 3. The scroll-down spotting of the h-index in the result list sorted by citedness in ADS

**APS** physics  
**Physical Review Online Archive**  
 AMERICAN PHYSICAL SOCIETY  
[Home](#) [Browse](#) [Search](#) [Members](#) [Subscriptions](#) [What's New](#) [Contact](#)

**Search Results (127 total)**

[Edit Your Search](#) / [New Search](#)

**Refine:** AND Abstract/Title  
**Sort By:** Most Recent Oldest First Most Cited

Showing results 1 - 100 of 127 total [Show](#)

1 2 [Next](#)

**Your Search**  
Author: j. e. hirsch

**Journals**  
[Phys. Rev. B](#) (105)  
[Phys. Rev. Lett.](#) (21)  
[Phys. Rev. A](#) (1)

**Category**

**Year**  
[1989](#) (14)  
[1983](#) (9)  
[1987](#) (9)  
[1988](#) (8)  
[1990](#) (7)  
[\(Show All Years\)](#)

**Icons**  
 PRL Editors' Suggestion  
 Free to Read

1. [Two-dimensional Hubbard model: Numerical simulation study](#)  
 J. E. Hirsch  
[Show Abstract](#)  
 Phys. Rev. B **31**, 4403 (1985)  
 Cited 313 times  
[PDF or Buy this Article](#)
2. [Attractive Interaction and Pairing in Fermion Systems with Strong On-Site Repulsion](#)  
 J. E. Hirsch  
[Show Abstract](#)  
 Phys. Rev. Lett. **54**, 1317 (1985)  
 Cited 256 times  
[PDF or Buy this Article](#)
3. [Spin Hall Effect](#)  
 J. E. Hirsch  
[Show Abstract](#)  
 Phys. Rev. Lett. **83**, 1834 (1999)  
 Cited 251 times

Figure 4. PROLA allows scroll-down spotting of the h-index in the result list which can be sorted by several criteria, including citation counts

On the other end of the spectrum, are the two largest databases of Ebsco. They do report citation counts in the master records but don't offer sort options. It would be possible to sort the results manually, but it is not worth the effort because so few records show citation counts. In the Ebsco MegaFILE database the h-index is 1, based on 4 references to the master records for 10 papers authored by J. E. Hirsch. In Ebsco Academic Search Premier the h-index is 0 because none of the 16 master records for papers of J.E. Hirsch show citation counts, as if none of them would have been cited by articles covered in this database. This is not so, because coming through the back door, and using the citation searching option, reveals such articles, but the citation matching algorithm does not see the match even when the references are accurate and would match master records.

Find: (((ZA "hirsch, j. e.") ) or ((ZA "hirsch, jorge") or (ZA "hirsch, jorge e."))) in Select a Field (optional) Search Clear

and Select a Field (optional)

and Select a Field (optional) Add Row

in: EBSCO MegaFILE ?

(Searching: *EBSCO MegaFILE*)

To store items added to the folder for a future session, [Sign In to M](#)

Refine Search Search History/Alerts Results

All Results: 1-10 of 10 Page: 1 Sort by: Date

See: All Results Academic Journals Magazines

**Narrow Results by Subject**

- [MAGNETISM](#)
- [MONTE Carlo method](#)
- [IONS](#)
- [COMPUTERS](#)
- [SCIENTISTS](#)
- [HILLS, Howard](#)
- [DOBSON, James -- Political & social views](#)
- [FERMI surfaces](#)
- [CITATION analysis](#)
- [CONDENSED matter](#)

1. [Does the h index have predictive power?](#) By: **Hirsch, J. E.** Proceedings of the National Academy of Sciences of the United States of America, 12/4/2007, Vol. 104 Issue 49, p19193-19198, 6p, 2 charts, 11 graphs; DOI: 10.1073/pnas.0707962104; (AN 27977726) [Cited References \(13\)](#)
2. [LETTERS.](#) By: O'Hearn, Katherine; Alterman, Eric; Minnery, Tom; Blumenthal, Max; De Brum, Tony; **Hirsch, Jorge**; Blank, David Eugene. Nation, 3/27/2006, Vol. 282 Issue 12, p2-30, 2p; (AN 20026603) [PDF Full Text](#) (227K)
3. [An index to quantify an individual's scientific research output.](#) By: **Hirsch, J. E.** Proceedings of the National Academy of Sciences of the United States of America, 11/15/2005, Vol. 102 Issue 46, p16569-16572, 4p; DOI: 10.1073/pnas.0507655102; (AN 19161021) [Cited References \(6\)](#) [Times Cited in this Database\(4\)](#)
4. [Superconductors as giant atoms: Qualitative aspects.](#) By: **Hirsch, J. E.** AIP Conference Proceedings, 2003, Vol. 695 Issue 1, p21-33, 13p; DOI: 10.1063/1.1639573; (AN 11805731) [Cited References \(42\)](#)

Figure 5. The h-index in Ebsco MegaFILE based on a single article cited 4 times in the database out of the 10 records for papers authored/co-authored by J.E. Hirsch

#### Database dimensions

These variable h-index values, and especially the absurdly low h-index values for J. E. Hirsch in the two Ebsco databases, offer the perfect example to illustrate the reasons for the significant differences, and the limitations -from the h-index perspective- of otherwise good quality, very large databases with cited reference enhanced records.

Of course, databases are not licensed with the purpose of determining the h-index of researchers. Ebsco keeps a low profile about the enhancement of many of its databases with cited references, and puts the emphasis on the enhancement of indexing/abstracting records with the full text of the source documents, and, indeed, this may be much more important for the majority of end-users. With that said, it is quite likely that h-indexes will be generated from several of the Ebsco databases that's why they are prominently discussed in this paper. Google Scholar, Scopus and Web of Science will have their own in-depth evaluation from the perspective of the h-index in the forthcoming issues, so they are discussed here only shortly and for context.

The completeness of a database referred to by Hirsch is a far more complex issue than the footnote in his seminal paper may suggest, i.e. that "the database used must be complete enough to cover the full period spanned by the individual's publications". The pre-

requisites are much broader. They include scope, size, retrospectivity, composition, breadth, and consistency of source coverage. These are relevant even for databases without cited references for comprehensive searches (Jacso, 1997), but for h-index calculation these requirements extend to the subset of the cited reference enhanced subset of the database. Not even the most accomplished scientists (except for information professionals specializing in bibliometric studies) may be fully aware of these traits of the databases, especially when it comes to the subset enhanced with cited references. The dimensions of that subset are usually the most enigmatic features of the databases.

### Size and scope

The size of the database is usually readily available for the searchers, but the size of the reference enhanced subset rarely is. For example, there is quite a variety of editions of the CINAHL database, and the catalog of Ebsco provides useful information about some of the content features, but not with regard to the cited reference enhanced subset of the different editions. All of them have such subset, usually between 18-20% of the entire database – except for CINAHL Select where 90% of the articles are available in full text, but merely 3.2% of the records are enhanced by cited references. This information is not readily available for the various CINAHL editions, and not even the librarians may know how to determine the number of records enhanced by cited references to get a feel about the comprehensiveness of the cited reference enhanced subset.

For many libraries, availability and searchability of full text may be much more important than the availability of records enhanced by cited references, and rightly so, even though the cited and citing references can significantly enrich the search experience and improve the results. After all, access to the full text is the meat for most users, and if one doesn't eat it with the most appropriate silverware it is less of a concern than the other way around, being equipped with the Martha Stewart endorsed utensils, but no meat to eat. With all that said, it is quite obvious from the results for the h-index test, that no matter how large is the Ebsco MegaFILE (33 million master records) or the Academic Search Premier database (14 million master records), they are not yet ready for calculating a reasonable h-index.

The size of the MegaFILE is practically the same as in Scopus (33.3 million records), and 10% larger than the 1980-2008 edition of Web of Science (30 million records). It is much smaller than the 1945-2008 edition of WOS, which has 39 million master records, but that is not the reason for its inadequacy for h-index calculation. The reason for that is that the size of the cited reference enhanced subset of about 735,000 records represent barely more the 2% of the database. In the smaller Academic Search Premier database the cited reference subset of 577,500 records represents 4% of the entire database. It is to be mentioned again, that in case of Web of Science, the library defines the time span (and thus the edition) to be licensed, in other cases the editions are predefined by the content providers, or it comes in just a single edition.

From the perspective of citation-based searching and the computing of the h-index the first main issue is the size of the cited reference enhanced subset of the databases not just the size of the entire database.

For comparison, in Scopus nearly 12.5 million records (37% of the full database of 33.3 million records) are enhanced by cited references. In the 1980-2008 edition of WoS this ratio is about 80%, yielding more than 23.6 million records enhanced with cited references. The 1945-2008 edition of WoS has 31 million cited reference enhanced records (80%) out of its 39 million records. In the edition of WoS covering 1996-2008, the number of records enhanced by cited references is about 12 million out of the nearly 15 million records, somewhat smaller than that subset in Scopus. Obviously, Ebsco is not in

the same league as Scopus or WoS, when it comes to cited reference enhanced subset, and that's why Ebsco does not advertise this feature widely.

One of the other reasons that makes MegaFILE and Academic Search Premier to perform so poorly for physicists is the poor coverage and the non-coverage of many of the top ranking physics journals. It is a double whammy for Hirsch who published most of his papers in journals not covered by these two Ebsco databases, and his papers got the majority of their citations from papers published in these very journals, not covered by these two databases of Ebsco.

The scope of the database is another factor to consider. When h-indexes are to be determined for researchers in a specific discipline, smaller databases with a focus on that discipline might serve the purpose very well, even though it may not be obvious from the name of the database. The open access Astrophysics Data System (ADS) << <http://adswww.harvard.edu> >> yielded very good results for my h-index tests for some of the physicists listed as most prominent in Hirsch's paper, even when their work had nothing to do with astronomy. The reason for this is that the database is as much for physics as for astronomy. Actually, ADS covers all physics areas not just astrophysics, in spite of its name, and Hirsch is mostly cited by articles in physics, except for his paper about the h-index which is much more widely cited in other science and social science journals than in physics journals. (ADS has more than 4 million records for physics papers, more than 1.5 million for astronomy, and 0.5 million are harvested from the arXiv repository).

If the scope of the cited reference enhanced database is very narrow but in that area its coverage is comprehensive and consistent, it can be used for determining the h-index for the researchers publishing in that disciplinary or sub-disciplinary area. A good example for this is the legendary SPIRES-HEP bibliographic database at Stanford, which was the first publicly accessible Web server in the U.S., set-up to offer open access to a bibliographic database.

Its design clearly shows the fingerprints of a librarian. It is Louise Addis, who was the highly interested and capable librarian at the Stanford Linear Accelerator Center (SLAC) when the SPIRES project was launched.

However, for the test of h-index of some of the most prominent physicists on Hirsch's lists, it could provide a realistic h-index only for Stephen Hawking. Actually, SPIRES-HEP is the database where Hawking gets his highest h-index,  $h=65$ , even above what the 1945-2008 WoS edition produces. The physicists on Hirsch list who work(ed) and published in other specialty areas of physics, had a very low h-index in SPIRES (Hirsch had  $h=2$ , and some Nobel laureates in physics got the same or  $h=1$ ) because the scope of the database is limited to high energy and particle physics, and it does a very good job at that, but not for physics in general.

### Retrospectivity

In discussing one of his test on determining the h-index for the Physics Nobel prize winners, Hirsch mentions that he used [for a calculation], the "first published paper year or 1955, the first year in the ISI database". Actually, it is the first year of coverage in the ISI database that his university licenses, but another university may have a larger edition of WoS, where the first year of coverage is 1945 or 1901.

Even the 10-year older edition of WoS can make considerable difference in determining the h-index for mature, well established researchers. The database edition used by Hirsch would not be fair for calculating the h-index for those Nobel laureates in physics, who were awarded in the early decades of the of the 20<sup>th</sup> century. No database can be fair to them because the papers of those Nobel laureates, and the thousands that cited them were written around the turn of the century, well before the start of coverage of most

databases. These are absent from the calculation of the h-index, except for the Century of Science edition of WoS, but even that edition covers only a small portion of their papers. The Chemical Abstracts, and PsycINFO databases also go back to the beginning of the 20<sup>th</sup> century, but those records are not enhanced by cited references.

Just think of Wilhelm Conrad Röntgen, the first Nobel Prize winner in the category in 1901. His most important papers (including the landmark paper “On a new kind of rays” published in *Science*) were written in the 1890s and received the majority of their citations before or around the turn of the century. Röntgen, Max Planck, Fermi and several other famous physicists would have an unrealistically low h-index from any databases (even from the Century of Science where their h-index is 2, 12 and 18, and Einstein achieves just 43). The shorter the time-span of the retrospective enhancement of records by cited references, the lower the h-index is - even for some of the most prominent researchers. This limitation in retrospective coverage is not navel gazing pondering about scientists who lived more than a century ago. In the mini case study of Hirsch of the Nobel Laureates in Physics of the past 20 years, the symptom of low h-index is there: the range in the 1955-2005 WoS edition is from 22 to 79, the average is 41, and the median is 35 – not an index value in physics that would indicate their Nobel-prize worthy research. It is not their fault. It is due to the limitations of the databases used.

To see what a difference even just a 10-year difference in retrospectivity can make in the h-index, I used the 1945-2008 and the 1955-2008 edition of WoS (Hirsch used the latter), and chose two physicists from the middle point of the two decades, Bertram N. Brockhouse and Clifford G. Shull who shared the 1994 Nobel prize.

For Brockhouse, the difference was just 1 point (h=33 in the former, and h=32 in the latter) between the two editions of WoS. However, in Shull’s case the h-index from the 1945-2008 edition was 35, and from the 1955-2008 it was only 26. The reason for this is that the latter did not include records for the source items and many of the citing items from the era (1946-1954) when Shull worked at what is known now as the Oak Ridge National Laboratory.

Further analysis shows that he wrote more than 40% of all of his papers during his years there. These earned more than 50% of the total citations his works received (or more precisely that could be attributed to him by the software). Using the WoS edition with a 10-year shorter time span does not include in the h-index generation process these cited and citing papers, and that’s the reason for the significantly lower h-index in Hirsch’s list for Shull.

Then again, the h-index computed from the 1955-2005 edition of WoS is still considerably higher than the h-index computed from other databases that can be used for h-index calculation in physics. In Scopus, Google Scholar, ADS, and PROLA the h-index for Shull were 12, 19, 18, and 19, respectively. Brockhouse’s h-index in Scopus (9) clearly shows the influence of its much shorter time span of records enhanced by cited references vis-à-vis Google Scholar (21), ADS (21) and PROLA (19). Depending on the target group or the individual whose “importance, significance, and broad impact” is to be measured, the span of the cited reference enhanced subset of the database is very important for fairness.

I discussed the dimensions of the cited reference enhanced subset of databases in a recent column (Jacso, 2007), but my focus was different there: the huge advantage of easily discovering items which cite or were cited by a few highly relevant papers on the users’ topic, irrespective of the terminological and morphological differences of the subject words. No library would license a cited reference enhanced database just for the sake of allowing a few users to determine the h-index of some researchers. It would be, however, increasingly important to inform the customers about this aspect of the database.

Thomson Scientific does not need to do this because whatever edition a library licenses, all the records include the cited references that appeared in the source documents. On the

other hand, in most other databases this type of enhancements of records with cited references goes back only to a few years.

The producers of the PsycINFO database deserve credit for providing the details about this aspect clearly, emphasizing that cited references have been added comprehensively only from 2001, and in the years before the enhancement is very selective. About 690,000 records are enhanced of the 2.4 million records in PsycINFO, and 86% of them are records for papers published after 2000. It may be used for relatively young researchers in psychology for determining their h-index, but for established researchers, the h-index produced from PsycINFO would not be fair. Scopus also goes out of its way to alert users that cited referenced were added to records only from 1996. (Actually there are some thousands records from earlier years which were also enhanced). In most of the cited reference enhanced databases this information is not mentioned.

For example, the Ebsco databases don't give a hint about the retrospectivity of the cited reference subset, but their small proportion (less than 5%) in the various editions (Premier, Complete, Alumni) of the Academic Search database might give a clue to the skilled users that such records are available only for the past few years, making their databases inappropriate for even gauging the "cumulative contributions" of scientists, as Hirsch meant to do it.

The public relation message of Chemical Abstracts seems to provide an easy to misunderstand statement, when its content description mentions that it has "links to cited references dating back to the beginning of the 20<sup>th</sup> century". Technically and legally it is true, but what the user would really need to know is the fact that cited references have been added only to records entered from 1997. Given its yearly addition of more than a million records, it undoubtedly offers the largest collection of reference enhanced records in chemistry from the past 11 years of the scientific literature, but for determining the h-index of chemists' whose contributions in the past 20-30 years have had the most influence, it is not the preferred resource.

### **Source base**

This criterion refers to the types of source documents covered by the database and enhanced by cited references. The PsycINFO database stands out from the crowd of traditional indexing/abstracting databases by covering a wide variety of primary sources beyond journals, such as monographs, collective works, book chapters, reference books, textbooks, conference proceedings, and dissertations. About 25% of the materials belong to one of these categories. From the reference enhanced subset perspective, the coverage is more limited because records for conference papers and dissertations have not been enhanced by cited references (yet), while 63,000 books and book chapters are already enhanced by cited references. This is true primarily for recent books and book chapters, recent meaning in this case the 21<sup>st</sup> century. Records for classic psychology books have not yet been enhanced, in spite of their importance in psychology.

In physics and in most of the hard sciences, books play a much smaller role in scientific communication than journal articles and conference papers. Still, book should not be ignored in the calculation of the h-index for those scientists who have put a lot of energy into writing books which became highly cited.

The most important move by Google Scholar since its launch was the autonomous indexing of cited references in books (harvested mostly from the Google Books service), and in conference proceedings. These improved significantly the quality of the source base, and in some cases, the h-index of researchers as I discuss in a section about stray and orphan references.

### **Journal base**

For decades, journals (and some other serial publications, such as Annual Reviews) have been the almost exclusive source documents processed by Thomson ISI for citation-based searching, and for calculating journal impact factors and other bibliometric measures. There is a well-defined set of scholarly journals in every discipline that can serve as a benchmark in judging the scope of journals and the breadth of their coverage by cited reference enhanced databases considered for computing a reasonably fair h-index. Their coverage must be also consistently comprehensive. These are usually not a problem in publishers' digital archives, but there are exceptions. In our professions, most of the journals of ALA have limited digital availability, and the same is true for the library science journals published by the Haworth Press. In information science John Wiley & Sons, does a very poor job in digitizing back issues of the *Annual Review of Information Science & Technology* (ARIST), as well as of the *Bulletin* and the *Proceedings of the American Society for Information Science* (the earlier name of the society). Its digitization abulia is just one of the reasons that make this publisher's digital archive totally inappropriate for h-index calculation of information scientists. Of the 42 volumes of ARIST, only 5 volumes from 2002 to 2006 are available digitally. For the *Bulletin* of the society, Wiley's Interscience database does not have in digital format the issues of its first 21 volumes, and the *Proceedings* of the society are included only for 2002-2006, missing 38 volumes. It needs further investigation to determine how incomplete the digital versions of its journals are in other disciplines. This lack of digital availability from the publisher is different from the situation when the library decides to license only a subset of the digital archives of journals (for reason of cost).

### **Consistency of coverage**

The coverage of journals in cited reference enhanced databases can be surprisingly uneven and spotty. Missing issues and volumes in indexing and abstracting records present a double whammy because in addition to not counting the articles for the productivity measure and their references for the citedness measure, the references of the papers citing articles in the missing issues don't have the master records as pegs to hang the citations on.

This is a problem with several journals in Ebsco Academic Search Premier which has only a few years of coverage for such journals as *Library and Information Science Research, Program, Aslib Proceedings, Journal of Internet Cataloging, Science & Technology Libraries, Online Information Review*, and have missing issues for many of them, such as *Knowledge Quest, Music Reference Service Quarterly, International Journal of Information Security*. It is to be noted that getting information about the consistency of the completeness of journal coverage is difficult and very time consuming even for a skilled information professional, and omissions can greatly influence the h-index of the authors of omitted papers, and of the journals themselves.

### **Software issues**

Currently, there are only a very few databases (including the publishers' digital archives) that have the essential tools for determining the h-index, let alone to do so efficiently. Disappointingly, the Dialog Information System, that was the first among the few to enhance its system for the handling of cited references in millions of records of ISI citation indexes a few decades ago under the most competent and compassionate management of the charismatic Roger Summit, has fallen behind the pack. A few years ago when PsycINFO records were partially enhanced by cited references there were failed efforts at first, and currently there are no options for searching by cited author, cited journal and cited year in PsycInfo on the Dialog platform. They are searchable as free text, such as

the abstracts, not as distinct elements as cited author, cited title word, cited year, cited journal. This is akin to eating soup with a fork.

No wonder there is no automatically generated information about the citedness of papers, let alone a sort option by this criterion, or automatic production of h-indexes in this implementation of PsycINFO. Neither was the software upgraded for the ISI citation databases to report automatically the citedness count of papers, let alone to sort the result lists by citedness count, or calculate the h-index.

The implementation of PsycINFO on Ovid and OCLC are far from ideal, and are inadequate for computing the h-index. Ebsco provides the citedness count in PsycINFO (and in many of its own databases), but does not offer sorting the result set by that information. CSA comes the closest to facilitate the calculation of the h-index by allowing the downloading of records directly to the spiffy RefWorks software, including the times-cited count. The set can be converted into a spreadsheet by the press of a button, and sorted by the citedness count in a swift process. (Ebsco also allows downloading records to RefWorks, but it does not send the citedness count data along with the bibliographic data, so from the perspective of calculating the h-index it is as useful as CPR to a dead person).

Beyond Web of Science and Scopus, only SPIRES, ADS, and PROLA offer adequate software tools for calculating the h-index. The valuable arXiv pre-print collection has many implementations, and it is one of the three components of ADS, which is the only host that provides citation counts, and allow sorting the sets by citedness.

Google Scholar does not rank the result list by citedness count systematically any more, which used to be the one and only sort possibility, neither does it offer the downloading of records. At least there are a number of third party utilities which can scrape the screen and even calculate the h-index. It is another question that these are at the mercy of Google, Inc. which can and does prevent access by these utilities to Google Scholar.

### Citation matching

The most critical question is the correct citation matching by the software. The variety of reference styles required by journals makes this process very error-prone. Instead of trying to describe the difficulties, I rather present an example – on the topic of citation matching from CiteSeer, one of the smartest autonomous citation matching system. What is shown here is just the tip of the iceberg. Papers that are cited by hundreds and thousands of times show much more variety than these 14 references that cite the same book in many different ways even in this simple case.

There are much more problematic documents from a referencing aspect, such as government documents, conference papers, large author groups. In case of journal articles, the various abbreviations and punctuations in journal titles, in volume, issue and page numbers bring up many additional problems. Author names with accented characters are a special challenge. Röntgen may appear also as Rontgen, Roentgen, or Rntgen when the accented character is swallowed by the software. German, Dutch, Italian, Spanish names with prefixes represent great challenge. Add to this double middle initials with or without punctuation and you get a picture about the difficulties of names even in simple Latin scripts. Author names with many adjacent consonants (as in Polish) are much more often misspelled than most of the Japanese names with the consonant-vowel pattern, such as Tanaka or Morimoto. Japanese, Chinese, and Korean names, however, are very problematic when they are transliterated to the same Latin character-string. The American Physical Society just announced an experiment to include the names of Chinese, Japanese, and Korean authors also in their own script to help distinguish their identity (Sprouse, 2008).

Hanna Pasula, Bhaskara Marthi, Brian Milch, Stuart Russell, and Ilya Shpitser. *Identity uncertainty and citation matching*. In Advances in Neural Information Processing (NIPS), 2002.

Hanna Pasula, Bhaskara Marthi, Brian Milch, Stuart Russell, and Ilya Shpitser. *Identity uncertainty and citation matching*. In Advances in Neural Information Processing (NIPS), 2003.

Hanna Pasula, Bhaskara Marthi, Brian Milch, Stuart Russell, and Ilya Shpitser. *Identity uncertainty and citation matching*. In Advances in Neural Information Processing (NIPS), 2003.

H. Pasula, B. Marthi, B. Milch, S. Russell, and I. Shpitser. *Identity uncertainty and citation matching*. In Advances in Neural Information Processing Systems 15. MIT Press. 2003.

H. Pasula, B. Marthi, B. Milch, S. Russell, and I. Shpitser. *Identity uncertainty and citation matching*. In NIPS 15. MIT Press, Cambridge, MA, 2003.

H. Pasula, B. Marthi, B. Milch, S. Russell, and I. Shpitser. *Identity Uncertainty and Citation Matching*. In S. Becker, S. Thrun, and K. Obermayer, editors, Advances in Neural Information Processing Systems 15. MIT Press, 2003.

H. Pasula, B. Marthi, B. Milch, S. Russell, and I. Shpitser. *Identity uncertainty and citation matching*. In NIPS, 2002.

H. Pasula, B. Marthi, B. Milch, S. Russell, and I. Shpitser. *Identity uncertainty and citation matching*. In Advances in Neural Info. Proc. Systems, 2002.

H. Pasula, B. Marthi, B. Milch, S. Russell, and I. Shpitser. *Identity uncertainty and citation matching*. In Proc. Advances in Neural Information Processing. MIT Press, 2003.

H. Pasula, B. Marthi, B. Milch, S. Russell, and I. Shpitser. *Identity uncertainty and citation matching*. In Advances in Neural Processing Systems 15, Vancouver, British Columbia, 2002. MIT Press.

H. Pasula, B. Marthi, B. Milch, S. Russell, and I. Shpitser. *Identity uncertainty and citation matching*, 2002.

H. Pasula, B. Marthi, B. Milch, S. Russell, and I. Shpitser. *Identity uncertainty and citation matching*. In Advances in Neural Information Processing (NIPS), 2003.

H. Pasula, B. Marthi, B. Milch, S. Russell, and I. Shpitser. *Identity uncertainty and citation matching*. In Advances in Neural Information Processing (NIPS), 2002.

H. Pasula, B. Marthi, B. Milch, S. Russell, and I. Shpitser. *Identity uncertainty and citation matching*. In NIPS, 2002.

Figure 6. Excerpt from the variety of references to a book with simple author names, no middle initials, prefixes and accented characters

Only SPIRES and Web of Science offer some help by allowing direct browsing of author names, and the former can also search automatically for name variants. Scopus has an indirect way to browse author and journal name variants to be discussed in my forthcoming review. The Institute of Physics does not even offer a search field for authors in its new interface, which is quite absurd. So is the new advanced search template of Windows Live Academic which shows a separate cell for 4 first and middle initials, but allows only one character, so you get 1,350 matches for Hirsch J. Hirsch J E was not among the first 100, not an encouraging start even for a plain vanilla search, let alone for determining a researcher's h-index.

Syntactically matching names don't necessarily give the green light for proceeding with the process. The set retrieved may include researchers with the name J. E. Hirsch, who are not the physicist, or are physicists, but not the one who worked at various campuses of the University of California, and started to publish in the mid-1970s. Their records must be excluded from the search. Obviously, the searcher must be familiar with the publication activity of the scientists whose h-indexes are to be determined, and must have good search skills to look up variant versions of the names, such as Jorge Hirsch, Jorge E. Hirsch, and do the limiting/filtering operations to exclude from the results set the hits for J. E. Hirsch, the audiologist, and J. E. Hirsch the surgeon, and decide if J. E. Hirsch of the National University of Buenos Aires should be retained in the set or not. Some software can help in this process.

Orphan and stray references

With all these oddities, another crucial issue is what happens with the references that cannot be matched. In some systems you may be able to browse or search them, and

manually add them to the master records which accrue the references. This is an arduous process (Jacso, 2008). No wonder that excepting the studies by Cronin and Meho (2006), Meho and Yang (2007), Oppenheim (2007), and a few others, most of the h-lists compilations known to me, relied only on the h-index values as reported automatically in the master records. But these do not include the stray references which do not match sufficiently for the software to add them to the master record. These may be in the thousands for some authors and are ignored in calculating their h-index or in spotting and consolidating them in the result list of cited reference searches.

In case of Google Scholar the problem is often the opposite. Its software finds a match even when there is not even a reference to the author, let alone to a specific work of the author in the purportedly citing document. These phantom citations can grossly inflate the hit counts in Google Scholar (Jacso, 2006). Even though the h-index is rather robust (Vanclay, 2007), and a few omissions for papers cited less often than the h-index would not cause significant changes, there is an additional important reference type whose handling may result in a considerable lower h-index than the researcher would deserve. I call these the orphan references, even though it is not an inherent feature, because they may become orphaned only in some cited reference enhanced databases, but not in others.

Using just the master records with the citation counts listed is always very convenient, but also often very unfair to the researchers whose h-index is to be determined. In addition to the stray cited references which cannot be matched against a master record to hang the citations on, there are the orphan references to books, chapters, conference papers and articles in journals not covered by the database as source documents, or are in issues omitted by the database producer for one reason or another. This is not strictly a matching issue because there are no master records for these items, so the references have nothing to be matched against. Inclusion of master records for books is probably the greatest asset of Google Scholar from a citation matching and h-index calculation perspective.

How the software handles the stray and orphan(ed) records is an important issue because these may significantly change the h-index of individuals who published highly cited books, as did S.W. Hawking in Hirsch's lists. In my tests, I found that he would have a 10-12% higher h-index if the citations received by his books from the articles in journals covered in WoS and Scopus could have been attached by the software to a master record, and thus taken into account in calculating his h-index. But books are not source documents in WoS, and in Scopus they became source documents -very selectively- only recently.

References given to books become orphans, because they can't be associated with and accrued through a master record, and thus are ignored in the automatic calculation of the h-index.

I realized what a difference this makes when I was preparing for a contribution to the Festschrift for the 75<sup>th</sup> birthday of F. W. Lancaster (Jacso, 2008), which was a noble opportunity to engage in the extra work needed to determine a reasonable h-index for such an important personality of the information industry. As that paper will be published later than this one, I don't want to reveal the extent of the increase of his h-index. Suffice it to say, that it was very much worth the arduous manual matching.

The forthcoming issues will review how the content and software features of Google Scholar, Scopus and Web of Science can meet the criteria for calculating a reasonable –if not perfect- h-index to assess scholarly productivity and impact.

## References

- Ashkanasy, N. M. (2007). "Playing the citations game", *Journal of Organizational Behavior*, Vol. 28 No. 6, pp. 643-645.
- Banks, M. G. (2006). "An extension of the Hirsch index: Indexing scientific topics and compounds", *Scientometrics*, Vol 69 No. 1, pp. 161-168.
- Bar-Ilan, J (2007). "Which h-index? – A comparison of WoS, Scopus and Google Scholar", *Scientometrics*, Vol. 74 No. 2, pp. 257-271.
- Berger, M. (2007). "The problematic ratings game in modern science.", *South African Journal of Science*, Vol. 103 No. 1-2, pp. 2-3.
- Braun, T., Glänzel, W., & Schubert, A. (2006). "A Hirsch-type index for journals", *Scientometrics*, Vol. 69 No. 1, pp. 169-173.
- Cronin, B and Meho, L. (2006). "Using the h-index to rank influential information scientists", *Journal of the American Society for Information Science and Technology*, Vol. 57 No. 9, pp.1275-1278.
- Egghe, L. (2006). "Theory and practise of the g-index". *Scientometrics*, Vol. 69 No. 1, pp. 131-152.
- Harzing, A and van der Wal, R. (2008) "A Google Scholar H-Index for Journals: A Better Metric to Measure Journal Impact in Economics & Business?", (submitted to the 2008 Academy of Management Annual Meeting August 8-13, 2008)
- Hirsch, J.E. (2005). "An index to quantify an individual's scientific research output". *Proceedings of the National Academies of Science*, Vol. 102 No. 46, pp. 16569-16572.
- Jacso, Peter.(2008). "Testing the calculation of a realistic h-index in Google Scholar, Scopus and Web of Science for F. W. Lancaster" (to appear in the special issue of *Library Trends*, Summer 2008).
- Jacso, P. (2007). "The dimensions of cited reference enhanced database subsets". *Online Information Review*, Vol. 31 No. 5, pp. 694-705.
- Jacso, P. (2006). "Deflated, inflated and phantom citation counts", *Online Information Review*, Vol.30, No. 3, 297-309
- Jacso, P. (1997). "Content Evaluation of Databases", *Annual Review of Information Science and Technology*, Vol 32: pp. 231-267
- Jin, B., Liang, L., Rousseau, R., & Egghe, L. (2007). "The R- and AR-indices: Complementing the h-index". *Chinese Science Bulletin*, Vol. 52 No. 6, pp. 855-863
- Meho, L. I. and Yang, K. (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*, Vol. 58 No. 13, pp. 2105-2125.
- Oppenheim, C. (2007). "Using the h-index to rank influential British researchers in information science and librarianship", *Journal of the American Society for Information Science and Technology*, Vol. 58, No. 21, pp. 297-301.
- Prathap, G. (2006). "Hirsch-type indices for ranking institutions' scientific research output". *Current Science*, Vol. 91, No 11, pp. 1439.
- Purvis, A. (2006). "The h index: Playing the numbers game", *Trends in Ecology and Evolution*, Vol. 21, No. 8, p. 422.
- Saad, G. (2006). "Exploring the h-index at the author and journal levels using bibliometric data of productive consumer scholars and business-related journals respectively", *Scientometrics*, Vol. 69, No 1, pp. 117-120.
- Schreiber, M. (2007a). A case study of the Hirsch index for 26 non-prominent physicists. *Annalen der Physik (Leipzig)*, Vol. 16 No. 9, pp. 640-652.
- Schreiber, M. (2007b). "Self-citation corrections for the Hirsch index". *Europhysics Letters*, Vol. 78 No. 3, Web-only publication at <http://dx.doi.org/10.1029/0295-5075/78/30002>
- Sprouse, G.D. (2008). "Which Wei Wang?" *Reviews of Modern Physics*, Vol. 80, p1.
- Vanclay, J. K. (2007). "On the robustness of the h-index". *Journal of the American Society for Information Science and Technology*, Vol. 58 No. 10, pp. 1547-1550.
- Vanclay, J. K. (2006). "Refining the h-index", *Scientist*, Vol. 20, No. 7, pp. 14-15
- Vinkler, P. (2007). "Eminence of scientists in the light of the h-index and other scientometric indicators". *Journal of Information Science*, Vol. 33, No. 4, pp. 481-491.