The H-Index: What Is It, How Do We Determine It, and How Can We Keep Up With It?

Timothy Ireland, Kathy MacDonald, and Peter Stirling

University of Waterloo, Canada

Much research has been undertaken about the h-index. What started out as a simple compound metric based on an individual's publications and citation counts has the potential to become increasingly complex and difficult to measure. We outline a simple but effective step-by-step process for creating comprehensive citation counts of an author's publications, and subsequently a more robust and accurate h-index based on results combined from multiple sources.

Introduction: What Is It?

In 2005, a physicist named Jorge E. Hirsch developed a simple premise in an effort to quantify the scientific output of an individual researcher.

> I propose the index h, defined as the number of papers with citation number ≤ h, as a useful index to calculate the scientific output of a researcher. (Hirsch, 2005, p. 16569)

To easily determine the h-index of a researcher, examine the number of times each paper has been cited and put them into descending order. Thus, if an individual has eight papers that have been cited 33, 30, 20, 15, seven, six, five, and four times, the individual's h-index would be six. The first paper, 33, gives a one—one paper has been cited at least once. The second paper gives a two—two papers have been cited at least twice. The third paper gives a three, and we continue all the way up to six with the sixth highest paper.

Tokar, A., Beurskens, M., Keuneke, S., Mahrt, M., Peters, I., Puschmann, C., van Treeck, T., & Weller, K. (Eds.). (2012). *Science and the Internet* (pp. 237-247). Düsseldorf: Düsseldorf University Press

The final two papers have no effect in this case because they have been cited less than six times.

Hirsch based his argument on the premise that the h-index is useful for comparing different researchers in similar fields. If the h-index is similar for the two people, their overall influence in the scientific field is similar, independent of the number of papers written or the number of overall citations. Likewise, an individual's h-index should increase linearly over time. It should be noted that the h-index is not the sole indicator of an individual's research impact and that its value varies between disciplines. Those disciplines not heavily invested in journal article publication and citation metrics as a measure of impact may find the h-index less useful.

The simplicity and ease of use / understanding led to the h-index metric being included in Thompson Reuters' ISI Web of Science (WoS) and Elsevier's Scopus "less than two years after its formation" (Zhang et al., 2011).

In the advancement and promotion process for faculty members at many Canadian universities, scholarly output is often an influential factor in determining tenure and promotion. Gathering citation counts for every article, conference proceedings, book chapters, and patents can be daunting. The added challenge is that the same article may be indexed in multiple databases in which some citing articles are the same between databases and some are unique.

WoS and Scopus collect and organize citation counts and can calculate an individual's h-index. Google Scholar does it via Google Scholar Citations. However, each source may determine a different value of the h-index for each individual. Sometimes the variation in the h-index between sources can be large. A person could take the highest citation or h-index counts from one of these databases and use them in tenure and promotion documentation or grant applications, but it may not be a full accounting of a person's h-index and may not be as accurate as it could be. Combining citation counts from various research databases gives a larger citation count and therefore a higher h-index. One can do this in such a way that it can be self-sustaining in terms of maintaining up-to-date citation counts and therefore h-index and provide documentation / proof of citation counts and h-index calculation. This paper outlines a step-by-step process on how to do this.

Brief Literature Review

There has been much research into the h-index and its variations. A recent review of the literature related to the publication, testing, and popularity of the h-index was summarized in an article by Zhang et al. (2011). The litera-

ture also points to the creation of variants of the h-index (such as the presence or absence of self-citations) to improve the metric. Likewise, the h-index has been applied to researchers from various fields and countries such as optometrists in Australia (Efron & Brennan, 2011), earth sciences (Mikki, 2010), psychology (Bador & Lafouge, 2011), chemical engineering (Prathap, 2011), medicine (Sanni & Zainab, 2011), and information science and library science (Levitt & Thelwall, 2009; Li et al., 2010; Meho & Yang, 2007; Oppenheim, 2007). There is extensive research into the calculation of an individual's h-index on different databases, as well as research into combining the results of databases (García-Pérez, 2010).

The research has suggested three main databases that should be used to determine the h-index of an individual scholar: WoS, Scopus, and Google Scholar (GS) Citations. Each provides an h-index based solely on the information it contains (content indexed), and as such is influenced by the strength and weakness of each database (Bar-Ilan, 2008). Substantial research has been done comparing the results of these three databases. Likewise, there has been considerable research into determining the accuracy of the h-index within each of these three databases. It has been established that an accurate h-index usually requires a compilation of multiple indexes (García-Pérez, 2010; Jacsó, 2008; Meho & Yang, 2007).

Question: How Do We Determine It?

Determining the H-Index Using Extensive and Exhaustive Searching

There is established methodology for creating de-duplicated and federated searches from databases to determine a more accurate h-index for an individual.

One study investigated 25 library and information science faculty members. The reported time for the project was over 3,300 hours (Meho & Yang, 2007), averaging about 132 hours per person. Overall, it would have taken a single person well over a year to complete the project. By the end of the project, the information would likely be out of date because some citations would have been added during the completion time. The study is comprehensive; however, most researchers would be unwilling or unable to dedicate that amount of time to determining their own h-index.

Could there be an easier, faster process that would not become outdated as soon as it was complete?

Challenges

A quick search using the established h-index databases (WoS, Scopus, and GS) reveals obvious differences in the results of the h-index for three researchers associated with the University of Waterloo (see Table 1). (The CVs of the individuals were provided to ensure accuracy.) If one of these researchers were to ask what their h-index is, what is the correct reply? Why are the numbers so different?

Table 1. A comparison of articles, total citations, and h-indices from WoS, Scopus, and GS for three researchers on June 15, 2012

	scholar 1	**scholar 2**	**scholar 3**
field of study	physics	psychology	psychology
first article published	2002	2003	1986
items in WoS	15	16	158
total citations	109	86	4032
WoS h-index	6	6	33
items in Scopus	19	23	146
total citations	106	247	3960
Scopus h-index	5	8	31
items in GS	43	28	195
total citations	177	544	7572
h-index in GS	8	11	40

In Table 2, comparing information in Gold Rush (i.e., a service that can be used to compare the holdings of various databases), we note that there is a substantial difference in the journals included within WoS and Scopus (GS information was unavailable). The Gold Rush search confirms that searching additional research database(s) adds the potential of finding overlooked unique journals not included in WoS or Scopus.

Within the current academic environment, research is no longer encompassed within a single subject area; it is more interdisciplinary. For researchers considered in Table 1, the addition of other databases in physics and psychology may yield additional citations. A search in multiple databases, indexing both the subject and specific journals, will increase the probability of finding additional citations. Ulrich's Web (i.e., a database that lists information about journals) provides information on which databases a specific journal may be indexed. Importing citations from multiple databases will also create duplicates, which will need to be managed carefully. The more suc-

cessful (as a measure of both publications and citations) an individual is, the more challenging this task becomes.

In some situations, author order and publication format are important, and there has been research into this (Levitt & Thelwall, 2009). More prestige may be given to authors who fall toward the front of the author listing. Likewise, what should be included as a citation may vary.

Table 2: Gold Rush comparison of WoS and Scopus on June 19, 2012

	Unique Journals	Similar Journals	Total Journals
WoS	2828	13053	15881
Scopus	14453	13053	27506

How Do We Keep Up With It?

Outline of Our Citation Tracking System

Developing research skills to determine individuals' scholarly impact may be essential to advancing an academic career (Hirsch, 2007). We now describe our methodology for citation tracking and determining a suitable h-index to be used for tenure and promotion applications, as well as tracking an individual's personal research influence. The process we outline can be set up easily for graduate students and faculty in the early years of their careers. Faculty members with many publications and a high h-index will find the process more time-consuming. The value of this citation tracking system is that it:

- creates a current list of all the academic output for an author, which may increase the accuracy of attributable citations;
- tracks author order and type of material published (for example, first author in a peer-reviewed journal);
- lists the author's citations in a transparent manner. This information can then be made publicly accessible;
- provides proof of combined citation counts and h-index calculation as each publication and all citing publications are collected and organized. This information can then be made publicly available;
- collects papers that cite their work so authors can monitor their impact and identify potential collaborators or competitors; and
- uses article citation alerts, automated e-mail, and RSS notification, making it easy to keep the database current.

Overview of the Process

The following steps outline our process for collecting and de-duplicating citations from multiple databases to calculate a more accurate h-index.

Step 1: Set Up the Author's Personal H-Index Tracking Database(s)

Using citation management software (CMS), set up two databases (the following process uses RefWorks). The first database (DB1) will include all of the academic output to be tracked. The second database (DB2) will contain all the citations to each of the works contained in DB1. An alternative option is to keep both academic output and citations of those works in the same RefWorks account using the folder level to collect author publications by publication type and the subfolder level to contain the citing works. For clarity, this paper will describe only the two database option.

Step 2: Populate DB1

Identify which research databases have citation index functionality and contain the author's publications. Some databases with this functionality include WoS, Scopus, PsycINFO, Social Science Research Network (SSRN), GS, etc. If the researcher is compiling a small number of publications, individual titles can be searched. However, a more useful strategy is to use the author finder search option. For example, in WoS, there is a tab with an author search / author finder option that refines by field, institution, and date range to narrow the author sets to a reviewable number. Scopus has a similar tab.

If searching a very common name results in too many author sets to sort efficiently, or if the name is not listed in the author sets, search for last name combined with each article title, or name and topic(s). If necessary, repeat the search using last name and the research topic relevant to the author's work.

A research profile can be set up with some research databases. This helps database indexers identify authors and thereby increases the accuracy of their citations. For example, WoS has a profile system called Researcher ID under the additional resources tab. Scholar Universe is another independent researcher profile system. These methods help databases credit authors with appropriate citations.

Once all the relevant articles are identified, the citations need to be imported into DB1.

Step 3: Populate DB2

DB2 contains all the relevant citations to all of the author's works in DB1. To track the citations, folders and subfolders will be used. The folders are labelled according to the publication types that are relevant for the creator's intended use. For example, most tenure and promotion documentation requires a tally of the citation counts for all refereed papers and separate citation counts for first-author papers compared with co-authored papers. The structure includes one folder each for all refereed first-author papers, for all refereed co-authored papers, for all conference proceedings, for book chapters, one for books, for patents, and for other publications such as government reports, white papers, thesis, dissertation, audio-visual modules, submissions to royal commissions, etc.

Subfolders can then be created within the publication type folders for each article, book, book chapter, conference proceeding, or other publication. For example, in the first-author folder, one should have a subfolder designated for each first-authored article. Each subfolder is labeled by the title of the publication. It may be beneficial to add the year if there are articles with similar titles published over a number of years. If useful, include a folder for posters, invited speaker presentations, patents—whatever is relevant for the discipline and stage of academic advancement. An efficient method of creating subfolders is to copypaste the title or the first part of the title of the article from DB1.

Step 4: Import the Citations into the Appropriate Subfolder in DB2

Return to an appropriate research database and do two things. First, identify the number of times the article, book, or conference proceeding has been cited in that database and export those citing articles into the designated subfolder for that particular article, book chapter, etc. Second, set up citation alerts for all publications in DB1 in each of the research databases used. If citation alerts are unavailable, saved searches may be available. Look for a "set alert," "e-alert," or "feed (RSS)" button. RSS feeds can be incorporated directly into a RefWorks account. The use of citation alerts allows for a relatively self-sustaining method for tracking.

If using GS, set up a direct export to RefWorks within Scholar Preferences or use RefGrab-It (i.e., a feature of RefWorks that allows citations on a Web page to be imported into RefWorks) to import multiple GS records. When reviewing records found in GS, consider what should be included in RefWorks. Consideration should be given to including or excluding certain

items, such as papers in other languages, advertisements, fact files, peer-reviewed trade publications, Web pages, posters, etc. GS does pick up citations to books, government publications, and white papers, etc. that may not appear in traditional databases, but nevertheless demonstrate the impact of a scholar's work.

Step 5: Move to the Next Appropriate Database and Repeat Steps 2 through 4

Step 6: Add Outstanding Items to DB1 and Add Originals if Desired

It is possible that not all of an author's academic output appears in a database, e.g., a non-governmental organization publication. These publications can be manually entered into RefWorks. RefWorks also offers the option to add attachments to citations. If desired, a text file, video clip, PDF, or other attachment may be added to the citation.

Step 7: Remove Unwanted Duplicates

One of the functionalities of RefWorks is the ability to identify duplicates automatically. DB1 should be free of all duplicates.

DB2 is a little more difficult in that duplicates are allowed in different subfolders. It is permissible to receive a citation for more than one item at the same time. For example, an article may cite several different articles from the same author with a single paper. Thus, duplicates are allowed, but not within the same subfolder.

Step 8: Harvest the Results

Determine final citation counts for each type of publication using the organize folders view, that is, all A1 (first-author), all CP (conference proceedings), etc. These numbers can then be transferred to tenure and promotion documentation, grant proposals, etc., or ranked to determine one's h-index.

Step 9: Keep the Database Up-to-date

Monitor e-mail or RSS alerts and add new articles citing the author's work to the appropriate citation subfolder to maintain a current record of citations for each publication, thus building the list of citations in preparation for an advancement process or other future use. Every time a new citation is added, the appropriate de-duplication process should be undertaken.

Step 10: Make Both DB1 and DB2 Public

DB1 and DB2 can be published or shared online, using RefShare, and made available for downloading into other CMS packages. This may increase both the frequency of an author's work being cited, and help ensure it is cited correctly. It also offers a transparent look into the compilation of an individual's h-index.

Conclusions and Future Research

To determine an individual's comprehensive h-index, searching multiple databases is required. A CMS streamlines the process immensely. Using a CMS creates a current list of all the academic output for an author. It builds a repository of publications with citations taken from one or multiple research databases. In doing so, authors have the opportunity to catch and correct any errors that database has made with respect to their publications, thus increasing the accuracy of citations to the author's work. The CMS contents can be shared by posting to a personal Web site or a departmental / university repository. The process allows for tracking various types of publications (first-authored papers, co-authored papers, patents, etc.) and their respective citation counts, which are required in most tenure and promotion documentation and some grant proposals. The author's citations can be presented in a transparent manner by making the citation database publicly accessible. The process also allows authors to collect papers that cite their work so authors can monitor their impact and identify potential collaborators or competitors. Through the use of article citation alerts, automated e-mail, and RSS notification, authors can keep their databases up-to-date and therefore have a current record of their citations counts and h-index.

There may be concerns about the creation of a system that offers different results than WoS or Scopus (or even GS). If an h-index from WoS and Scopus are seen as authoritative, it is not a stretch to look at combining the results of these two databases (and removing duplicate records). Our research has shown that authors with an h-index of 31 in WoS and 33 in Scopus ended up with a calculated index of 34 using our method.

Likewise, the inclusion or exclusion of a cited reference type will quite obviously affect the h-index of a researcher. The outstanding challenge is determining exactly what should be included and what should be excluded. Further research and discussion about the types of publications to be included in an h-index calculation is required.

Acknowledgements

Our thanks are due to Sandra Keys and Annie Bélanger for assistance editing this paper.

References

BADOR, P., & LAFOUGE, T. (2011). Comparative analysis between impact factor and h-index for psychiatry journals. *Canadian Journal of Information and Library Science*, *35*(2), pp. 109-121.

BAR-ILAN, J. (2008). Which h-index? A comparison of WoS, Scopus and Google Scholar. *Scientometrics*, *74*(2), pp. 257-271.

EFRON, N., & BRENNAN, N. A. (2011). Citation analysis of Australia-trained optometrists. *Clinical and Experimental Optometry*, *94*(6), pp. 600-605.

GARCÍA-PÉREZ, M. A. (2010). Accuracy and completeness of publication and citation records in the Web of Science, PsycINFO, and Google Scholar: A case study for the computation of h-indices in psychology. *Journal of the American Society for Information Science & Technology*, *61*(10), pp. 2070-2085.

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), pp. 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*, *104*(49), pp. 19193-19198.

JACSÓ, P. (2008). Testing the calculation of a realistic h-index in Google Scholar, Scopus, and Web of Science for F. W. Lancaster. *Library Trends*, *56*(4), pp. 784-815.

LEVITT, J. M., & THELWALL, M. (2009). The most highly cited library and information science articles: Interdisciplinarity, first authors and citation patterns. *Scientometrics*, *78*(1), pp. 45-67.

LI, J., SANDERSON, M., WILLETT, P., NORRIS, M., & OPPENHEIM, C. (2010). Ranking of library and information science researchers: Comparison of data sources for correlating citation data, and expert judgments. *Journal of Informetrics*, *4*(4), pp. 554-563.

MEHO, L. I., & 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*, *58*(13), pp. 2105-2125.

MIKKI, S. (2010). Comparing Google Scholar and ISI Web of Science for earth sciences. *Scientometrics*, *82*(2), pp. 321-331.

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 & Technology*, *58*(2), pp. 297-301.

PRATHAP, G. (2011). Correlation between h-index, Eigenfactor™ and article influence™ of Chemical Engineering journals. *Current Science*, *100*(9), p. 1276.

SANNI, S. A., & ZAINAB, A. N. (2011). Evaluating the influence of a medical journal using Google Scholar. *Learned Publishing*, *24*(2), pp. 145-154.

ZHANG, L., THIJS, B., & GLÄNZEL, W. (2011). The diffusion of h-related literature. *Journal of Informetrics*, *5*(4), pp. 583-593.