Scientometrics (2021) 126:6137–6145
https://doi.org/10.1007/s11192-020-03699-9

The *h*-index formalism

Leo Egghe[1] · Ronald Rousseau[2,3]

Received: 22 July 2020 / Published online: 14 September 2020

Abstract
This article provides an overview of the development of the *h*-index formalism. We begin with the original formulation as provided by Hirsch and move on to the latest versions. In this we show how the *h*-index formalism has evolved over time. Lesser known versions, in particular the continuous version of the *h*-index is brought to the front. We also discuss the Prathap–Kosmulski–Schubert successive *h*-indices. As announced in the title we focus on the *h*-index formalism, omitting generalizations, such as the *g*-index, and applications in research assessment.

Keywords *h*-Index · Generalized *h*-index · Discrete and continuous case · Successive *h*-indices · Higher-order *h*-indices

Introduction

Traditionally serious doubt has been raised about the reliability of bibliometrics as applied to individuals or small groups (King 1987). Narin (1976, p.160) already pointed out several pitfalls when applying citation analysis to individual scientists. This point has recently been stressed again by Moed (2020), who referred to calculating indicators at the level of an individual and stated that claiming that such indicators by themselves measure an individual's performance, is a case of false precision. Yet, in 2005 this idea has been challenged by Hirsch (2005) who introduced a new indicator based on publications and citations, and aiming particularly at the evaluation of individuals. Although this indicator, known as the *h*-index, has several disadvantages when applied in a research evaluation setting, it took bibliometrics by storm and became one of the most popular indicators. As Hirsch' original publication inspired hundreds of scientists, soon several reviews were published, see e.g., Rousseau (2008), Alonso et al. (2009), Egghe (2010) and Schubert and Schubert (2019).

✉ Ronald Rousseau
ronald.rousseau@uantwerpen.be; ronald.rousseau@kuleuven.be

Leo Egghe
leo.egghe@uhasselt.be

1 University of Hasselt, Hasselt, Belgium

2 Faculty of Social Sciences, University of Antwerp, 2020 Antwerp, Belgium

3 Centre for R&D Monitoring (ECOOM) and Department of MSI, KU Leuven, 3000 Leuven, Belgium

Recent books too devoted considerable attention to the *h*-index and its generalizations (Todeschini and Baccini 2016; Vitanov 2016; Rousseau et al. 2018).

It should be observed that the *h*-index is not obtained as a simple mathematical operation on the data, such as adding, multiplying or dividing, applied to numbers of publications or received citations, but is obtained as the result of a process. Indeed data must be ranked and then a comparison must be made to obtain the *h*-index (see further for the complete definition). In the next sections we will ignore the *h*-index as an indicator for research evaluation but instead study the mathematical formalism.

Formalizing: discrete version

In this section, we show how the same discrete formalism has been defined in different situations and with different purposes.

The *h*-index for authors

Slightly adapting the original formulation as published by Hirsch (2005), the *h*-index for authors is defined as follows (Rousseau et al. 2018):

> 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 (the exact order does not matter). Then the h-index of scientist S 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′ h-index is h if the number h is the largest natural number (representing a rank) such that the first h publications received each at least h citations.

Hirsch (2005) further noted that ideally, one should eliminate self-citations. We also remark that, as formulated by Hirsch, a scientist's *h*-index is a total career index. Yet, it is just a small step to include a publication window, e.g. only considering publications over the latest 20 years, and/or a citation window, e.g. only considering citations received at most 3 years after the publication year (van Raan 2006; Liang and Rousseau 2009).

The *h*-index for sources other than scientists

Hirsch considered scientists, their publications and their citations. It is now a simple step to consider other sources of publications, such as journals. This is precisely what has been proposed by Braun et al. (2005, 2006). These authors used a 1-year publication and a five-year citation window.

Already in 2006 Liu Zeyuan (to whom this article is dedicated) and colleagues from Dalian University of Technology, studied the *h*-index for Chinese journals in library and information science, and in management science, based on data from the CSSCI (Chinese Social Sciences Citation Index) (Jiang et al. 2006). They compared the *h*-index of these journals, with the relative *h*-index for journals (the *h*-index divided by the number of publications), the *g*-index and the journal impact factor (Garfield and Sher 1963; Egghe 2006a, b). The authors warned that these indicators should be used with caution if applied for journal evaluation.

In the same year, Rousseau (2006a) showed how the *h*-index of JASIS changed for publications between 1991 and 2000 and a—variable—citation window ending in 2005. This is one of the first time-dependent studies of an h-index.

Besides journals, one may apply the same scheme to institutions, countries, etc… as sources of publications. An early example for countries is provided in Csajbók et al. (2007), while Prathap (2006) considered institutions. Prathap's idea is discussed further on in the section on successive indices.

Other source-item relations

Besides publications as sources and citations as items, 'produced' by these sources, the *h*-index scheme may also be applied to other source-item relations. Probably the first to do this was Banks (2006) who performed a topic search (via TS=), in particular about chemical compounds, and ranked the corresponding publications decreasingly according to received citations. The *h*-index algorithm could then be applied to these lists. The corresponding *h*-index provides a 'hotness' value for the corresponding topic or chemical compound, especially when a relatively short publication window is applied. Another early case is the example of books and their loans in a university library (Liu and Rousseau 2007, 2009).

A special case occurs when one uses years as sources and publications as items, or years as sources and received citations (in that year) as items. Moreover, one may in a similar vein consider years as sources, and citations received by articles published in that year as items. All these cases are referred to as year-based *h*-indices (Mahbuba and Rousseau 2013), see also Rousseau et al. (2018).

Finally, in the case of so-called successive *h*-indices (see further), *h*-indices themselves play the role of items, replacing numbers of citations in Hirsch' original formulation.

In all these cases the same formalism is used as that proposed by Hirsch. Consequently, all resulting indices should be referred to as *h*-indices.

h-Indices for infinitely long sequences

Another small step, but still using the same formalism consists of using infinitely long sequences instead of a finite list of observed data. We recall the definition from Egghe and Rousseau (2019a).

> Let $(\mathbf{R}^{+})^{\infty}$ be the positive cone of all infinite sequences with non-negative real values and let $X = (x_r)_{r=1,2,\ldots}$ be a decreasing array in $(\mathbf{R}^{+})^{\infty}$. The h-index of X, denoted h(X), is the largest natural number such that the first h coordinates have each at least a value h. We observe that if all components of a decreasing array X are strictly smaller than 1, then $h(X) = 0$.

As infinitely long sequences do not occur in practical cases, this generalization only applies in theoretical work. Yet, any finite data array can be considered as an infinitely one, by adding infinitely many zeros. Again, there is no reason to refer to the corresponding *h*-index by another name.

The van Eck–Waltman h_θ-index

Before formulating this *h*-type index, we first recall the definition of the *h*-index as introduced by Hirsch:

> A scientist S′ h-index is h if the number h is the largest natural number such that the first h publications received each at least h citations.

Now van Eck and Waltman (2008) noted that there is nothing special about h articles each receiving at least h citations. Hence they proposed the following modification:

> given $\theta > 0$, the h_θ-index of scientist S is h_θ if h_θ is the largest natural number such that the first h_θ publications received each at least θh_θ citations.

We denote the h_θ-index of scientist *S*, by $h_\theta(S)$. If $X = \left(x_s\right)_s$ is the array of received citations for each of *S*′ publications, ranked in decreasing order of received citations, then, given $\theta > 0$, the definition of the h_θ-index of scientist *S* can be reformulated as follows

> given $\theta > 0$, the h_θ-index of scientist S is h_θ if the number h_θ is the largest index such that $x_{h_\theta} \geq \theta h_\theta$.

Moreover, if $x_1 < \theta$ then $h_\theta(S) = 0$. Although the van Eck–Waltman h_θ-index is a rather small modification of the original version, it must be observed that in this way the single notion of an *h*-index has been extended to an infinitely long one-parameter family of *h*-type indices with θ as parameter. Their importance will become clear in the next section.

Continuous versions

In this section, we move from the concrete to the abstract mathematical situation. Indices are not anymore calculated for finite or infinite sequences, but for continuous functions.

The *h*-index for a continuous curve

Let $Z(r)$ be a decreasing continuous curve defined on the positive real line. Then the abscissa of the intersection of the line $y = r$ and $y = Z(r)$ is called the *h*-index of $Z(r)$, in short $h(Z)$. As far as we know this model was first used in Egghe and Rousseau (2006) while calculating the (continuous) *h*-index in a Lotkaian situation. In Egghe and Rousseau (2019b) we were able to derive a new geometric interpretation of the *h*-index of *Z* which makes use of the Lorenz curve of *Z*, providing another illustration of the usefulness of the continuous case. If one considers the discrete case and connects points with line segments leading to a continuous curve, then the corresponding *h*-index was already introduced in Rousseau (2006b). It is nowadays known as the interpolated *h*-index (Rousseau et al. 2018, p. 213). The relation between the continuous *h*-index and the interpolated one is illustrated in Fig. 1.

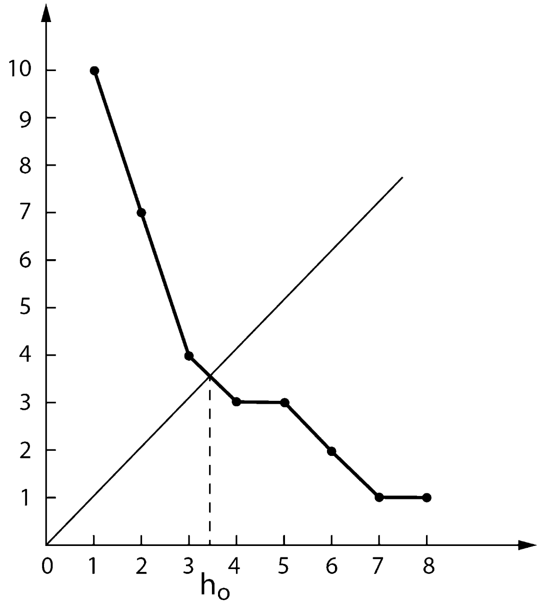

Fig. 1 A discrete finite array and how to obtain its interpolated *h*-index (indicated as h_0)

The continuous van Eck–Waltman *h*-index

In Egghe and Rousseau (2019c) we introduced the analog of the van Eck–Waltman *h*-index as follows. Given the positive, non-constant, decreasing, and differentiable function $Z(r)$ defined on the interval $[0, T]$, we define the functions $h_\theta(Z)$, $\theta > 0$, as the unique solution of the equation:

$$Z\left(h_\theta(Z)\right) = \theta h_\theta(Z). \tag{1}$$

We further note that given $Z(r)$ and θ, Eq. (1) only applies for $Z(T) \leq \theta T$. We add that for the definition itself the function $Z(r)$ does not have to be differentiable. Moreover, the definition is also valid for a constant function. Indeed, it is easy to see that for

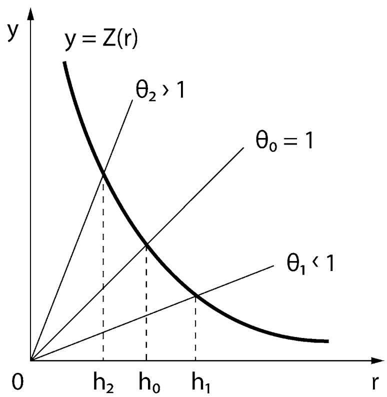

Fig. 2 Construction to obtain the general $h_\theta(Z)$-index; h_0 is the standard (continuous) *h*-index

$Z(r) = a$ (constant), $h_\theta(Z) = a/\theta$. The $h_\theta(Z)$-index is illustrated in Fig. 2. In this figure the index of h on the r-axis is just a number corresponding to the index of θ.

It is clear that for these functions Z, $h_\theta(Z)$ is strictly decreasing in θ, indeed: $\theta_1 < \theta_2 \Rightarrow h_{\theta_1}(Z) > h_{\theta_2}(Z)$. We further see that, if $Z_1 \leq Z_2$ (with same domain) then, for fixed θ: $h_\theta\left(Z_1\right) \leq h_\theta\left(Z_2\right)$, if these generalized h-indices exist.

The h_θ-index plays an essential role in solving the following problem: find the value a in $[0, T]$, such that the L^2-distance between the functions Z and φ_a reaches a minimum (Egghe and Rousseau 2019c), where φ_a denotes the function which is equal to a (a natural number at least equal to 1) times the characteristic function of $[0, a]$.

The case of not necessarily decreasing functions

As the continuous h_θ-index is obtained as the abscissa of the intersection of a straight line and a decreasing function $Z(r)$, one may ask if it is really necessary that this function is decreasing. In Egghe and Rousseau (2020a, b) the authors answered this question in the negative and introduced the general h_θ-index. They moreover established a relationship with polar coordinates. Concretely, for the continuous function $Z(r)$, they obtain a general h_θ-index, denoted $h_\theta(Z)$, if one of the following equivalent conditions holds:

1. For every $r \in \mathcal{D}(Z)$, there exist a unique $\theta > 0$ such that $Z(r) = \theta r$;
2. $\mathcal{R}(Z(r)/r) = \mathcal{D}(h(Z))$;
3. The graph $y = Z(r)$ can, for every $r \in \mathcal{D}(Z)$ be described in polar coordinates as a polar function (not just a relation) $\rho(\varphi)$ with domain a subset of $]0, \pi/2[$.

In the preceding lines, the symbols $\mathcal{D}$ and $\mathcal{R}$ stand respectively for the domain and the range of a function. Now, given $\theta > 0$, the unique r-value for which $Z(r) = \theta r$, is defined as $h_\theta(Z)$. It is then shown in Egghe and Rousseau (2020a, b) that these generalized h-indices may lead to a description of a Cartesian equation of a function in polar coordinates. Concretely, under the conditions mentioned above, $y = Z(r)$, for $r \in \mathcal{D}(Z)$, has the following polar form: $\rho_Z(\varphi) = h_\theta(Z) \cdot \sqrt{1 + \theta^2}$, with $\theta = \tan(\varphi)$. This construction is illustrated in Fig. 3.

Further comments on this construction can be found in Smolinsky (2020). Once an h-index for not necessarily decreasing functions has been established one may define a discrete version for not necessarily decreasing arrays, somewhat similar to the interpolated h-index, but we will not go into this.

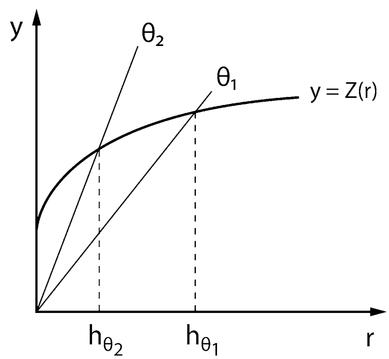

Fig. 3 Construction of the general h-index for an increasing continuous function $Z(r)$

A generalization of the *h*-index: the Prathap–Kosmulski–Schubert successive indices

In this section, we study an interesting 'higher dimensional' case, of which the history is less known, and which is not discussed in the latest review by Schubert and Schubert (2019).

Hirsch (2005) already considered the calculation of an *h*-index for a group of individuals. In this respect, he made the following observation. The overall *h*-index of a group will generally be larger than that of each of the members of the group but smaller than the sum of the individual *h*-indices, because some of the papers that contribute to each individual's *h*-value will no longer contribute to the group's *h*.

Prathap's *h*-indices

Prathap (2006) proposed two *h*-type indices to evaluate an institution's output. The first one, denoted h_1, is just the *h*-index of all publications of an institute (again using a given publication and citation window). Such an *h*-index has been called a global *h*-index (Egghe 2008).

The second *h*-index, denoted as h_2, is calculated as follows. One first calculates the *h*-index of each scientist working at the institute. This list of *h*-indices can now be ranked from highest to lowest, and a new *h*-index, h_2, is obtained as an *h*-index of *h*-indices.

We observe that Prathap actually considered three types of *h*-indices: *h*-indices of scientists, *h*-indices of the union of all publications of scientists belonging to an institute (h_1), and then a second-order *h*-index (h_2) as an *h*-index of *h*-indices of scientists.

Some years later, Rousseau et al. (2010) considered the ratio h_2/h_1 as a structural indicator, related to the stability of the research performed at an institute or university. Here, the term stability is used in the sense of not depending on a small group of scientists that could easily move to another university or institute.

Kosmulski's *h*-indices

In the same year as Prathap and completely independently, Kosmulski (2006) discussed the *h*-index for scientists and, sidestepping the obvious h_1, immediately moved on to Prathap's h_2 (which he denoted by *i*) for an institute. He further calculated h_2 without self-citations.

Schubert's successive *h*-indices

The next year Schubert (2007) generalized this setting leading to so-called successive *h*-indices. In a similar way as an institute consists of scientists, he considered—as an example—a country 'consisting' of publishers, and publishers 'consisting' of journals. He then calculates the *h*-index of each journal (as in the second section, and then calculates a second-order *h*-index for publishers (similar to Prathap's h_2). Finally, he moves one step further to calculate a third-order *h*-index for a country as an *h*-index of the country's publishers' *h*-indices. Egghe (2008) modeled successive *h*-indices in a Lotkaian framework.

Conclusion

It has been convincingly shown (Bouyssou and Marchant 2011; Waltman and van Eck 2012) that in research evaluation applications the *h*-index may behave contrary to what he is designed for. Like many other bibliometric indicators it is only "probably approximately correct (PAC)" (Rousseau 2016) and may be said to only have heuristic value for evaluation purposes. Yet, in our opinion, this should not deter scientists to study the logical and mathematical properties of the *h*-index formalism itself, and search for applications other than in research evaluation.

In this contribution, we focused on the *h*-index formalism, but in recent work (Egghe and Rousseau 2019a, b, c, 2020a, b) we also included the *g*-index formalism and several other generalizations.

References

Alonso, S., Cabrerizo, F. J., Herrera-Viedma, E., & Herrera, F. (2009). *H*-Index: A review focused in its variants, computation and standardization for different scientific fields. *Journal of Informetrics, 3*(4), 273–289.
Banks, M. G. (2006). An extension of the Hirsch index: Indexing scientific topics and compounds. *Scientometrics, 69*(1), 161–168.
Bouyssou, D., & Marchant, T. (2011). Ranking scientists and departments in a consistent manner. *Journal of the American Society for Information Science and Technology, 62*(9), 1761–1769.
Braun, T., Glänzel, W., & Schubert, A. (2005). A Hirsch-type index for journals. *The Scientist, 19*(22), 8.
Braun, T., Glänzel, W., & Schubert, A. (2006). A Hirsch-type index for journals. *Scientometrics, 69*(1), 169–173.
Csajbók, E., Berhidi, A., Vasas, L., & Schubert, A. (2007). Hirsch-index for countries based on Essential Science Indicators data. *Scientometrics, 73*(1), 91–117.
Egghe, L. (2006a). How to improve the *h*-index. *The Scientist, 20*(3), 14.
Egghe, L. (2006b). An improvement of the *h*-index: the *g*-index. *ISSI Newsletter, 2*(1), 8–9.
Egghe, L. (2008). Modelling successive *h*-indices. *Scientometrics, 77*(3), 377–387.
Egghe, L. (2010). The Hirsch-index and related impact measures. In B. Cronin (Ed.), *Annual review in information science and technology* (Vol. 44, pp. 65–114). Medford, NJ: Information Today Inc.
Egghe, L., & Rousseau, R. (2006). An informetric model for the Hirsch-index. *Scientometrics, 69*(1), 121–129.
Egghe, L., & Rousseau, R. (2019a). Infinite sequences and their *h*-type indices. *Journal of Informetrics, 13*(1), 291–298.
Egghe, L., & Rousseau, R. (2019b). A geometric relation between the *h*-index and the Lorenz curve. *Scientometrics, 119*(2), 1281–1284.
Egghe, L., & Rousseau, R. (2019c). Solution by step functions of a minimum problem in $L2[0, T]$, using generalized *h*- and *g*-indices. *Journal of Informetrics, 13*(3), 785–792.
Egghe, L., & Rousseau, R. (2020a). *h*-Type indices, partial sums and the majorization order. *Quantitative Science Studies, 1*(1), 320–330.
Egghe, L., & Rousseau, R. (2020b). Polar coordinates and generalized *h*-type indices. *Journal of Informetrics, 14*(2), 101024.
Garfield, E., & Sher, I. H. (1963). New factors in the evaluation of scientific literature through citation indexing. *American Documentation, 14*(3), 195–201.
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.
Jiang, C. L., Liu, Z. Y., & Liang, Y. X. (2006). Introduction to *H*-index and *G*-index for journal academic impact evaluation. *Library and Information Service, 50*(12), 63–65. **(in Chinese)**.
King, J. (1987). A review of bibliometric and other science indicators and their role in research evaluation. *Journal of Information Science, 13*(5), 261–276.
Kosmulski, M. (2006). "I"—A bibliometric index. *Forum Akademickie, 11,* 31. **(In Polish)**.

Liang, L. M., & Rousseau, R. (2009). A general approach to citation analysis and an *h*-index based on the standard impact factor framework. In B. Larsen & J. Leta (Eds.), *Proceedings of ISSI 2009—The 12th international conference on scientometrics and informetrics* (pp. 143–153). Rio de Janeiro: BIREME/PAHO/WHO and Federal University of Rio de Janeiro.

Liu, Y. X., & Rousseau, R. (2007). Hirsch-type indices and library management: The case of Tongji University Library. In D. Torres-Salinas and H. F. Moed (Eds.), *Proceedings of ISSI 2007, 11th international conference of the International Society for Scientometrics and Informetrics* (pp. 514–522). Madrid: CINDOC-CSIC.

Liu, Y. X., & Rousseau, R. (2009). Properties of Hirsch-type indices: The case of library classification categories. *Scientometrics, 79*(2), 235–248.

Mahbuba, D., & Rousseau, R. (2013). Year-based *h*-type indicators. *Scientometrics, 96*(3), 785–797.

Moed, H. F. (2020). Appropriate use of metrics in research assessment of autonomous academic institutions. *Scholarly Assessment Reports, 2*(1), 1.

Narin, F. (1976). *Evaluative bibliometrics*. Cherry Hill, NJ: Computer Horizons Inc.

Prathap, G. (2006). Hirsch-type indices for ranking institutions' scientific research output. *Current Science, 91*(11), 1439.

Rousseau, R. (2006a). A case study: Evolution of JASIS' *h*-index. *Science Focus*, *1*(1), 16–17. **(in Chinese)**. An English translation can be found at http://hdl.handle.net/10760/7616.

Rousseau, R. (2006b). Simple models and the corresponding *h*- and *g*-index. Unpublished. Available at http://hdl.handle.net/10760/7501.

Rousseau, R. (2008). Reflections on recent developments of the *h*-index and *h*-type indices. *COLLNET Journal of Scientometrics and Information Management, 2*(1), 1–8.

Rousseau, R. (2016). Citation data as proxy for quality or scientific influence are at best PAC (probably approximately correct). *Journal of the Association for Information Science and Technology, 67*(12), 3092–3094.

Rousseau, R., Egghe, L., & Guns, R. (2018). *Becoming metric-wise. A bibliometric guide for researchers*. Oxford: Chandos (Elsevier).

Rousseau, R., Yang, L. Y., & Yue, T. (2010). A discussion of Prathap's h_2-index for institutional evaluation with an application in the field of HIV infection and therapy. *Journal of Informetrics, 4*(2), 175–184.

Schubert, A. (2007). Successive *h*-indices. *Scientometrics, 70*(1), 201–205.

Schubert, A., & Schubert, G. (2019). All along the *h*-index-related literature: A guided tour. In W. Glänzel, H. F. Moed, U. Schmoch, & M. Thelwall (Eds.), *Springer handbook of science and technology indicators* (pp. 301–334). Cham: Springer.

Smolinsky, L. (2020). Two theorems on "Polar coordinates and generalized *h*-type indices". *Journal of Informetrics, 14*(3), 101045.

Todeschini, R., & Baccini, A. (2016). *Handbook of bibliometric indicators. Quantitative tools for studying and evaluating research*. Weinheim: Wiley-VCH.

van Eck, N. J., & Waltman, L. (2008). Generalizing the *h*- and *g*-indices. *Journal of Informetrics, 2*(4), 263–271.

van Raan, A. F. J. (2006). Comparison of the Hirsch-index with standard bibliometric indicators and with peer judgment for 147 chemistry research groups. *Scientometrics, 67*(3), 491–502.

Vitanov, N. K. (2016). *Science dynamics and research production. Indicators, indexes, statistical laws and mathematical models*. Cham: Springer.

Waltman, L., & van Eck, N. J. (2012). The inconsistency of the *h*-index. *Journal of the American Society for Information Science and Technology, 63*(2), 406–415.