Scientometrics (2020) 122:1303–1322
https://doi.org/10.1007/s11192-020-03364-1

Exploring the limitations of the h-index and h-type indexes in measuring the research performance of authors

Jingda Ding[1] · **Chao Liu**[1] · **Goodluck Asobenie Kandonga**[1]

Received: 24 March 2019 / Published online: 25 January 2020

Abstract
With the introduction of an increasing number of evaluation indexes, researchers have begun to pay attention to the limitations of such indexes in research evaluation, understanding which to avoid misusing and making evaluation more scientific and reasonable. Analysing the principles of the h-index, g-index, AR-index, p-index, integrated impact indicator (I3), and academic trace, this paper explores their limitations in measuring the research performance of authors from the perspectives of consistency, the degree of discrimination, and the statistical relationship between the values of indicators and the number of publications and citations. There are some interesting findings. These six indicators are highly consistent, and they are all more susceptible to the number of publications than to the frequency of citations. Among them, the h-index has the lowest degree of discrimination, followed by the g-index, I3, AR-index, p-index, and academic trace. The g-index ignores papers and citations other than the g-core. Moreover, compared to the h-index, the accumulation of citations makes it easier for the g-index to be equal to the number of papers published by an author, and once its value equals the number of papers, subsequent citations received by these papers will no longer contribute to the growth of the g-index unless the author publishes a new paper. Additionally, the AR-index ignores the h-tail papers and citations, which underestimates the impact of many researchers. Moreover, the p-index is insensitive to highly cited papers. Furthermore, the I3 is very vulnerable to the influence of the extremums in a data set. Finally, we propose considerations and suggestions for the research performance evaluation of authors.

Keywords h-index · h-type indexes · Academic performance evaluation · Comparative research · Limitations of indicators

Introduction

Research performance evaluation always plays an important role in scientific development, providing benchmarks for recruitment, promotion, funding, and rewards. To make research evaluation scientific and reasonable, various bibliometric indexes have been successively

✉ Jingda Ding
djdhyn@126.com

[1] Department of Library, Information and Archives, Shanghai University, Shanghai, China

proposed. However, each indicator has inherent limitations in measuring authors' research performance (Agarwal et al. 2016). These limitations make it challenging to find the appropriate indicator in the practice of research evaluation and even lead to the abuse of bibliometric indicators, or they make the measurement results inconsistent with the actual situation. Therefore, it is necessary to explore the limitations of bibliometric indexes to improve research performance evaluation; in particular, it is important to correctly understand how to use such indexes and how to interpret the measurement results. Based on theoretical and empirical analysis, this paper discusses the limitations of the h-index and h-type indexes such as the g-index, AR-index, p-index, integrated impact indicator (I3), and academic trace, each of which improves a defect of the h-index, which is described in detail in the Literature Review and Methodology. Hirsch proposed the h-index (Hirsch 2005) based on the hypothesis that the number of citations a scientist receives reflects the relevance of his/her work more than the number of papers that he/she publishes. Egghe (2006) proposed the g-index in an attempt to modify or avoid some limitations of the h-index. Jin et al. (2007) put forward the AR-index to measure the h-core's citation intensity and also takes the age of publications into account. Subsequently, conducting an in-depth study of the h-index, Prathap (2010) proposed the p-index, which reveals the arithmetic relationship between the h-index and the number of papers and citations. The discussion on the standardization of bibliometric reference counting led to the emergence of the I3 (Gingras and Larivière 2011), which can remedy some of the shortcomings of previous indexes to an extent. Then, the academic trace was proposed in 2017; this indicator combines the h-index and I3, taking into account the entire distribution of the citation curve (Ye and Leydesdorff 2014). Although the follow-up indicators noted above were proposed to improve research evaluation, each indicator has inherent limitations. Therefore, it is necessary to explore and identify their limitations. The research objectives of this paper are as follows:

- To analyse and discuss the inherent defects and deficiencies of the six indexes in measuring the research performance of authors based on their design principles.
- To compare the differences among the six indicators in measuring the research performance of a same author sample set and to reveal and summarize their limitations based on theoretical analysis.
- To propose some suggestions with regard to using these six indexes to measure the research performance of authors.

Literature review

The initial performance evaluation of authors was based on the number of published papers, which essentially measured productivity (Agarwal et al. 2016). With the establishment of Garfield's scientific citation index (SCI) system, the frequency of citation was the proxy index of academic influence. However, both measures can reflect only one aspect of research performance evaluation. Later, Hirsch (2005) proposed the h-index, which has attracted increasing attention from researchers because of its robustness and simplicity. Ball (2005) also affirmed the validity of the h-index in *Nature*. Although there are different opinions, research on the h-index and its use in practice are in full swing.

Although the h-index was initially proposed as a proxy for the number of citations, it combined the number of papers and citations. It has also been widely used in the performance evaluation of authors. Hirsch himself (2005) recommended the guidelines of $h > 12$

at a minimum for the promotion of a physicist at a leading university to associate professor and $h > 18$ for promotion to full professor. Moreover, in the mandates imposed by the recent reform of Italian higher education, the recruitment of associate and full professors must take place through national competitions that are open only to those who exceed the threshold of certain bibliometric indicators, including the h-index (Abramo et al. 2013a). Although it is widely used, the h-index has some limitations, such as the incomparability between disciplines, the fact that low-cited and zero-cited papers are ignored, and its insensitivity to papers with a high number of citations (Costas and Bordons 2008; Bornmann et al. 2008; Alonso et al. 2010). Therefore, many h-type indexes have been proposed to overcome the shortcomings of the h-index.

Many indexes have been proposed to eliminate the impact of author cooperation and disciplinary differences. Batista et al. (2006) divided the h-index by the average number of authors in h papers and designed the h_I index. Sidiropoulos et al. used the h_f index to obtain the normalized h-index, which aimed to achieve a direct comparison between different subjects (Sidiropoulos et al. 2007; Iglesias and Pecharromán 2007). Given the shortcomings of the h_I index, Schreiber (2008b) proposed the h_m index, which eliminated the defect of the h_I index of excessively reducing the influence of large-scale cooperative papers and increasing the influence of individually authored papers. Hirsch (2010) proposed the ħ index, which considers the influence of collaborators in measuring the research performance of authors. The hα index proposed by Hirsch (2019) was a useful complement to the h-index of a scientist to quantify his/her scientific achievement, rectifying an inherent drawback of the h-index, that is, the inability to distinguish between authors with different co-authorship patterns.

Additionally, many scholars have made efforts to compensate for the lack of sensitivity of the h-index to highly cited papers. The g-index proposed by Egghe (2006) changed the citations of a single paper in the h-index to the cumulative citations of papers, which increased the sensitivity of the h-index to highly cited papers. Jin (2006), Jin et al. (2007) proposed the A-index, R-index, and AR-index, focusing on the concept of the h-core, and assessed the impact of high-level papers in a natural way. In particular, based on the R-index, AR-index goes one step further and takes the age of publications into account. This allows for an index that can actually increase and decrease over time. Zhang (2009b) also proposed the w-index, which assigns different weights to the citations received by different articles, thus improving the sensitivity of the h-index to highly cited papers.

In addition, Prathap (2010) proposed the p-index and revealed the arithmetic relationship between the h-index and the number of papers and citations. He thought that the p-index could be interpreted as a factor of prestige or prominence to compensate for the lack of sensitivity and discrimination of the h-index. Moreover, the discussion on bibliometric citation count standardization led to the emergence of the I3 (Gingras and Larivière 2011), which uses percentiles to classify categories according to the citation frequencies of papers (Wagner and Leydesdorff 2012). Bornmann (2013) showed how to conduct a meaningful analysis of the percentiles in evaluation studies, suggesting that bibliometric evaluation should pay attention not only to the percentage distribution of publications but also to influential publications. Similar indicators have been applied in the science and engineering indicators of the US National Science Council (Bornmann et al. 2008). Subsequently, Ye et al. proposed the academic trace (Ye and Leydesdorff 2014; Ye et al. 2017), which combines the I3 and h-index to overcome the defect of each.

Bornmann (2014) observed that research on the h-index yielded approximately 50 kinds of h-type indexes. Additionally, Bornmann et al. (2008) divided the h-index and nine h-type indexes into two categories, the impact core and the output core, pointing out that

peer evaluation could be predicted better using the former than using the latter. Bornmann et al. (2011) also found that there was a high correlation between the h-index and 37 kinds of h-type indexes, including the g-index, A-index, AR-index, and H_I index. Barnes (2017) reviewed the debate on the h-index in bibliometrics, pointing out that although the h-index is prevalent in higher education as a decision-making tool, there are still some basic questions about its accuracy in measuring research performance.

Though the literature review, we think the basic questions of h-index are as follows. The h-index is insensitive to highly cited papers; the h-index value only increases and does not decrease over time; the degree of discrimination for the h-index is low; the h-index ignores most citations and papers outside the h-core. Therefore, we choose a representative index to improve the defects in each of the four aspects of the h-index. The g-index increases the sensitivity of the h-index to highly cited papers; the AR-index takes the age of publications into account, which allows for an index that can actually increase and decrease over time; the p-index compensates for the lack of discrimination of the h-index; the I3 uses percentiles to divide the papers into different categories, which consider the whole distribution of publications and citations. Finally, the academic trace is a representative comprehensive index; it is necessary to explore its rationality in measuring scholars' research performance. Therefore, we choose these five h-type indexes: the g-index, AR-index, p-index, I3, and academic trace, as well as the h-index, to analyse their advantages, limitations, and conditions of applicability to provide some suggestions for the research evaluation of authors.

Methodology

Theoretical analysis of each index

An author has index h if the h of his/her Np papers has at least h citations each and the other (Np-h) papers have $\leq$h citations each (Hirsch 2005). In other words, an author's h-index indicates that at most h papers are cited at least h times and that the h papers are the most productive core of the author's output, known as the Hirsch core or h-core (Rousseau 2006; Burrell 2007). To rank papers from high to low according to their citation frequency, when the citation frequency of a paper is greater than or equal to its ranking order, the value of the h-index is the same as the maximum ordinal number. According to Hirsch's model, a researcher's total citations increase over time in a linear fashion. He also assumed that if the researchers did not publish any more papers, then the slope of h-index tends to be stable with the growth of time, rather than showing a discontinuous change state. However, Liang (2006) confirmed that only one of the physicists in Hirsch's model demonstrated the expected linear increase. Therefore, if Hirsch's model does not, in fact, correspond to reality, there is no reason to expect that a comparison between researchers in terms of their h-index scores will lead to meaningful results (Barnes 2017). Moreover, an increasing number of bibliometricians are now convinced that the construction of the h-index is inherently arbitrary (Abramo et al. 2013b). Although it has simplicity and robustness, the h-index also has some limitations. It pays attention to only h-core papers, ignores most papers with a low citation frequency, and lacks sensitivity to highly cited papers (Zhang 2013). If a paper is already in the h-core, then the citations that it receives will no longer contribute to the growth of the h-index. Additionally, if two or more authors have an h-index with the same value, then we need additional indicators to distinguish and evaluate them. It only increases but never decreases, which is unfair for the researchers who have

just stepped into the academic field. The growth of h-index is a dynamic process. At the beginning of a scientist's research career, due to the limitation of the number of papers and citations, the h-index shows a slow-growth state; after a period of time, the number of papers and citations have a certain amount of accumulation, and the h-index shows a rapid growth state; with the change of time, the number of papers and citations have been growing, then the h-index grows all the time; after it reaches the maturity stage, because the number of papers is no longer growing, and the number of citations will continue to grow, so the h-index shows a slow growth until it remains unchanged.

To compensate for the insensitivity of the h-index to highly cited papers, Egghe (2006) proposed the g-index. Among the many h-type indexes, the g-index is mostly discussed (Tol 2008; Schreiber 2008a; Woeginger 2009; Schreiber 2009). According to the definition of the g-index, a set of papers has g-index g if g is the highest rank, such that the top g papers altogether have at least g^2 citations. The collection of papers in the g-index is called the g-core, and theoretical and empirical studies have shown that the g-index value is close to the average number of citations of the g-core papers (Schreiber 2010b). Similar to the h-index, it is also necessary to sort papers by citation frequency from high to low; then, the value of the g-index is the same as the maximum ordinal number when the total number of citations is greater than or equal to the square of the ranking order. The g-index pays attention to the cumulative citation frequency of papers, which improves its sensitivity to highly cited papers, but it ignores papers and their citations if the papers are not included in the g-core (Abramo et al. 2013a). Moreover, compared to the h-index, the accumulation of citations makes it easier for the g-index to be equal to the number of papers published by an author. Once the value is equal to the number of papers, the citations received by these papers will no longer contribute to the growth of the g-index unless the author publishes a new paper. In addition, the value of the g-index is an integer, which tends to cause multiple authors to have the same g-index, making it impossible to distinguish their research performance.

The construction process of AR-index needs to start from A-index. In 2006, Jin proposed A-index to measure the h-core's citation intensity, which achieves the same goal as the g-index, namely correcting for the fact that the original h-index does not take the exact number of citations of articles included in the h-core into account. This index is simply defined as the average number of citations received by the publications included in the h-core. But it also brings a problem that the better scientist is "punished" for having a higher h-index, as the A-index involves a division by h (Jin et al. 2007). Therefore, in 2007, Jin et al. (2007) put forward R-index, which is defined as the square root of the sum of citations received in the h-core. Taking the square root has the advantage of leading to indicator values which are not very high and of the same dimension as the A-index. In order to overcome the problem that the h-index may never decrease and that scientists may, so to speak, 'rest on their laurels', Jin et al. (2007) proposed AR-index based on the R-index. The formula of AR-index is as follows:

$$\mathrm{AR} = \sqrt{\sum_{j=1}^{h} \frac{\mathrm{cit}_j}{a_j}} \tag{1}$$

In the formula (1), a_j denotes the age of article j, cit_j denotes the number of citations article j received. If there are several publications with exactly h citations, then we include the most recent ones in the h-core. It not only takes the actual number of citations into account but also makes use of the age of the publications. In this way, the

h-index is complemented by an index that can decrease actually. In addition, it covers all citations received by the publications in h-core, so that the changes of citations in h-core papers may also have an effect on the value of AR-index. But it also has a disadvantage that only considers the publications in h-core, ignoring the rest of publications and citations even though it is considerable, which may underestimate the impact of many researchers.

The p-index reveals the arithmetic relationship between the h-index and the number of papers and citations. The formula of the p-index is as follows:

$$P = (C(C/N))^{1/3} \tag{2}$$

In formula (2), N represents the number of papers, and C is the frequency of citations received by the papers. The calculation of the p-index is straightforward, involving only the number of papers and the number of citations. Its value is usually a decimal, which largely prevents different authors from having the same p-index value. However, the p-index ignores the citation distribution. As long as the total number of papers and the total number of citations are the same, the value of the p-index will be the same, which to some extent weakens the role of highly cited papers in evaluating an author's research performance.

The discussion on the standardization of bibliometric reference counting has led to the emergence of the I3 (Gingras and Larivière 2011). Consider a set A, a reference set S containing all elements in A, hence $A \subseteq S$, and a function g from S to the positive real numbers, leading to the multiset g(S). Note that we consider g(S) as a multiset as we consider the images g(s), s in S, as separate entities (even if their values are the same). Our study situation is the case that A consists of a set of articles of an author, the set S consists of all articles of all authors our study selected in which the set A is contained, and a function g which maps an article to the number of citations it has received (and there may be several articles with the same number of citations). Now a rule is given which subdivides the set S into K disjoint classes, based on the values of the function g (and this independent of A). If a document belongs to class k then it receives a score x_k, where x_k does not depend on A. A standard situation is a case that there are 100 percentile classes. In the case of percentiles articles belonging to the top 1% receive a score of 100, those belonging to the top 2% (and not to the top 1%) receive a score of 99, and so on. The formula of the I3 is as follows (Leydesdorff and Bornmann 2011; Rousseau and Ye 2012):

$$I3(A) = \sum_{k=1}^{k} x_k * A(k) \tag{3}$$

If we take $A = S$, divide S $(= A)$ into two classes, namely the h-core and the h-tail, and give those articles in the h-core a score of 1 and those in the h-tail a score of 0 then I3(A) is exactly equal to the h-index of A. This shows that the h-index is, at least formally, a special case of the I3 score. Although "the h-index can be written in such a way that it formally looks like an I3 score, it is not an I3 score. The reason is that the scores x_k and the classes may not depend on the set A." (Rousseau and Ye 2012). From the definition, the I3 not only considers the whole citation distribution but also improves the sensitivity to highly cited papers by giving them higher weights. Although the I3 is adaptable to different data sets, changes in the extreme values in a data set seriously affect its measurement result. The size of extreme values is the basis for classification. If the extreme values are too large, most papers will be divided into categories with low

scores due to the insufficient citation frequency, resulting in a low I3 value for most authors, which also reduces the degree of discrimination of the I3 to an extent. Therefore, the I3 is not suitable for data sets with excessive citation frequency extremums.

The academic trace is an indicator that integrates the h-index and I3, and it not only overcomes the shortcoming of the h-index with regard to ignoring papers outside the h-core but also overcomes the defect of the I3 with regard to being easily affected by the extreme values in a data set. The reasoning process of the academic trace is as follows. In general, if we rank papers from high to low according to their citations, we can obtain a Citation-Publication (C-P) rank distribution—the citation curve—as shown in Fig. 1 (Ye and Leydesdorff 2014). There are three sections relevant to the h-index: the h-core, the h-tail, and the un-cited (zero citations) papers (Pz). Furthermore, Zhang (2009a) called the area above the h-core a representation of "excess citations", that is, citations that are gathered but do not further contribute to the h-value.

Combined with the idea of the I3, three vectors are used to mark the distribution of all publications and citations: the publication parameters (X), the citation parameters (Y) and the different parameters between high and zero citations (Z).

$$X = (X_1, X_2, X_3) = (P_c^2/P, P_t^2/P, P_z^2/P) \tag{4}$$

$$Y = (Y_1, Y_2, Y_3) = (C_c^2/C, C_t^2/C, C_e^2/C) \tag{5}$$

$$Z = (Z_1, Z_2, Z_3) = (Y_1 - X_1, Y_2 - X_2, Y_3 - X_3) \tag{6}$$

where $P = P_c + P_t + P_z$, $C = C_c + C_t + C_e$. P_c, P_t, P_z represent the number of publications in h-core, h-tail and the number of publications received zero citations, respectively. C_c, C_t, C_e represent the number of citations in h-core, h-tail, and h-excess, respectively. So, the vector X and Y indicate the distributions of publications and citations in the h-core, h-tail and the uncited as well as excess area, respectively. The terms of Z can be appreciated as the fraction of citations minus the fractions of publications so that Z is a set of meaningful indicators, where Z_3 is a complex indicator because one considers the excess citations as possible compensation for the uncited publications. The fraction of uncited publications contributes negatively to Z_3, but this can be compensated for by the fraction of excess citations in a set.

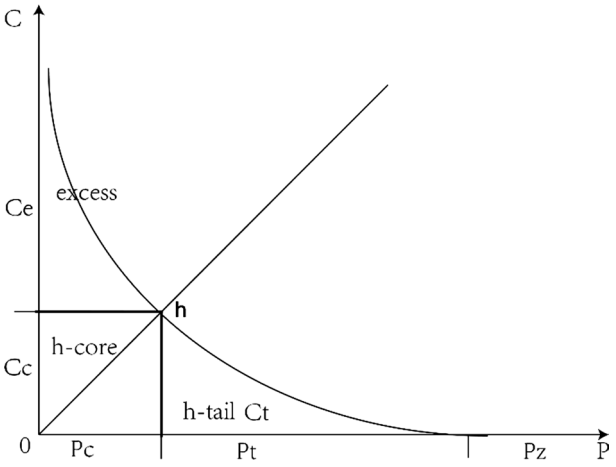


Fig. 1 The C-P rank distribution curve

The academic trace is the trace of the matrix composed of these three vectors, and its formulas are as follows:

$$T = tr(V) = Y_1 + X_2 + Z_3 = C_c^2/C + P_t^2/P + (C_e^2/C - P_z^2/P) \tag{7}$$

$$T = tr(V) = \frac{h^4 + (C_h - h^2)^2}{C} + \frac{(P - h - P_z)^2 - P_z^2}{P} \tag{8}$$

The academic trace is a mathematical result that follows naturally from the core-tail framework of the h-index when combined with the idea of relative frequencies used for I3. There are five parameters in formula (8): P is the number of publications, P_z is the number of zero-cited papers, C is the number of citations, C_{h} is the number of citations in the h-core, $C_{\mathrm{h}} = C_{\mathrm{c}} + C_{\mathrm{e}}$, and h is the h-index value of a set of papers. When the T value is positive, the higher it is, the better the academic performance of the author. In contrast, if T the value is negative, then an author's academic performance will be reduced. The academic trace considers the overall distribution of publications and their citations, providing more comprehensive and abundant measurement information. However, like most indicators, the academic trace, which is based on the citation frequency, h-index and number of papers, depends heavily on the accumulation of time, which is suitable for scholars of a certain academic age but not for young researchers. At this time, instant evaluation indicators such as the Faculty of 1000 (F1000) can be an excellent complement. Additionally, the calculation of the academic trace is more complex, which may also affect its practical application in practice.

Empirical research on the six indexes

In this paper, we selected 106 chemistry scholars from Harvard University as the research object. First, in the Web of Science (WOS) database, the subjects were limited to chemistry, the institutions were limited to Harvard University, and time was limited to 2009–2018. A total of 6, 979 articles were retrieved, and their bibliographic information was downloaded. Second, according to the information in the address field, we deleted the authors and pieces of literature not belonging to Harvard University and retained 5874 works of literature. We then selected 106 authors who have published more than 15 papers in the past 10 years. The reason for setting 15 as the threshold is to ensure the activity of authors in research and to prevent the six index values of authors publishing fewer than 15 papers from being too small to be compared. Third, we downloaded the bibliographic information and citation information of the 106 authors. Finally, the h-index, g-index, AR-index, p-index, I3, and academic trace values of each author were calculated.

Based on the theoretical analysis of these six indicators, we further discuss the limitations of each indicator in measuring the authors' research performance through empirical analysis. This paper used a series of statistical methods, such as correlation analysis and linear fitting, to explore the degree of discrimination of each index, the consistency of the measurement results, the relationships among indexes and the number of papers and citations, and the limitations of each index.

Table 1 The value of each index for the top 20 authors in descending order of the h-index

Authors	Papers	Citations	h-index	g-index	AR-index	p-index	I3	Academic trace
Weissleder, Ralph	438	36,261	98	176	51.055	144.256	1079	9542.203
Weitz, David A.	425	26,166	85	146	44.944	117.227	863	6107.816
Khademhosseini, Ali	432	24,373	85	136	44.613	111.202	816	5114.529
Whitesides, George M.	399	32,369	82	172	50.061	137.963	983	11,617.772
Capasso, Federico	357	21,394	67	141	47.280	108.636	737	8034.308
Lieber, Charles M.	148	20,662	65	143	42.027	107.309	571	11,920.058
Mooney, David J.	246	17,435	65	129	42.971	142.351	561	7000.023
Suo, Zhigang	211	14,002	64	114	37.026	97.581	446	4583.620
Farokhzad, Omid C.	99	23,385	63	99	49.946	176.772	587	15,382.624
Mahadevan, L.	273	11,949	61	103	30.830	107.613	460	3666.329
Ingber, Donald E.	177	14,852	61	121	39.055	80.569	452	6762.087
Toner, Mehmet	167	16,265	61	127	40.572	116.573	473	7750.417
Gray, Nathanael S.	159	15,521	58	124	40.295	114.854	446	7494.392
Hamblin, Michael R.	229	13,301	56	107	32.176	91.758	439	3721.023
Walsh, Christopher T.	207	11,554	53	102	28.609	86.397	385	3551.493
Clardy, Jon	175	8583	51	88	27.247	74.946	309	2700.666
Xie, X. Sunney	137	10,145	50	100	31.745	90.906	326	4890.756
Aizenberg, Joanna	175	8530	49	90	31.824	74.637	309	3145.254
Wagner, Gerhard	167	8356	49	88	25.771	74.776	294	2832.014
Liu, David R.	127	11,286	47	106	28.833	100.098	336	5857.499

Table 2 The min, max, mean (m), standard deviation (sd), and median (mdn) of the six indicators for the 106 authors

Indicators	Min	Max	*m*	SD	mdn
h-index	7	98	30.66	19.369	25.5
g-index	12	176	57	37.582	48
AR-index	4.763	51.055	20.087	11.463	17.752
p-index	10.308	176.772	54.466	33.131	47.987
I3	16	1079	185.05	212.035	99
Academic trace	62.815	15,382.624	2150.224	2766.255	1158.604

Results

Descriptive statistics

Table 1 shows the number of papers and citations and the h-index, g-index, AR-index, p-index, I3, and the academic trace values of the top 20 authors in descending order of the h-index. Table 2 shows the min, max, mean (m), standard deviation (sd), and median (mdn) of the six indicators for the 106 authors.

As shown in Table 1, the h-index, g-index and AR-index values are no higher than the number of papers published by the author, while the I3 values are the opposite, that is, no lower than the number of papers published by the author. This result indicates that the growth of the h-index, g-index and AR-index will be limited by the number of papers published by authors. Table 2 shows that the span of the AR-index is 4.763–51.055, which is the smallest among the six indexes, and the span of the academic trace is 62.815–15382.624, which is the largest, with the span of the other four indicators falling in between.

The h-index, g-index, and I3 are ranked from large to small according to their value, followed by the I3, g-index, and h-index. For each author, the p-index value is no lower than that of the h-index, and for most authors, the p-index value is higher than that of the g-index but less than that of the I3. AR-index is developed on the basis of R-index considering the publication time of article. Although the value of R-index is always greater than or equal to the value of h-index, AR-index is not necessarily. In this study, the academic trace value is significantly larger than the values of the other indicators, but this fact is not always the case. For example, when an author publishes only one paper and is not cited, the academic trace value is -1, while the value of the other four indicators is non-negative.

Degree of discrimination for the six indicators

The degree of discrimination is an important criterion for evaluating the quality of indicators. In this section, we rank the degree of discrimination of the six indicators by comparing the number of authors with the same value in each index. The number of author groups with the same value and the numerical range of authors in each group for each index are shown in Table 3.

For the h-index, 33 author groups have the same value, and each group includes 2–7 authors. An author with a smaller h-index value has a higher probability of repeating. Among the 106 authors, the h-index values of Liang, Steven H., McCarthy, Michael C., Saikin, Semion K., Zheng Shao-Liang, Chorev, Michael Lee, David, Y. W., and Reus, William F. are the same, 14. Therefore, it is difficult to distinguish their academic performance based on the h-index. Additionally, there are significant differences in the number of papers published and the number of citations received by these seven authors, and their g-index, p-index, I3 and academic trace values are not the same. Therefore, the h-index has the lowest degree of discrimination among the six indexes. The g-index is slightly better than the h-index in terms of the degree of discrimination. There are 26 groups of authors having

Table 3 The number of author groups with the same value and the numerical range of authors in each group with the same value for each index

	h-index	g-index	AR-index	I3	p-index	Academic trace
The number of author groups with the same value	33	26	0	20	0	0
The numerical range of authors in each group	2–7	2–4	0	2–3	0	0

the same g-index value, with each group including 2–4 authors. For example, Shum Ho Cheung, Kim Shin-Hyun, Madix, Robert J., and Myers, Andrew G. have the same g-index, 33. Additionally, the degree of discrimination of the I3 is slightly better than that of the g-index; 20 author groups have the same I3 value, with each group including 2–3 authors. No two authors share the same AR-index value, p-index value or the same academic trace value. Because their calculation results usually contain decimals, they largely avoid repeating the same value. In addition, Table 2 shows that the span of the academic trace is larger than that of the p-index, and the span of the p-index is larger than that of the AR-index, which helps prevent the repetition of the same value and improves their degree of discrimination. Therefore, if we rank these six indexes from high to low according to the degree of discrimination, the order is the academic trace, p-index, AR-index, I3, g-index, and h-index.

Consistency of the six indicators

Correlation analysis can measure the consistency between indicators. The higher the correlation coefficient is, the stronger the consistency between the indicators, and vice versa. SPSS 21.0 was used to analyse the correlations between these six indicators, and the correlation coefficients are shown in Table 4. These six indicators are highly correlated with each other, which indicates that the measurement results of the six indicators have high consistency. The correlation between the academic trace and the other five indexes is the lowest, indicating that the academic trace has a relatively independent value and significance in measuring authors' academic influence.

The relationship of each indicator with the number of publications and citations

The number of publications and the number of citations are the two fundamental dimensions of academic quantitative measurement. Not only are they the source of other evaluation indexes, but they are also the cornerstone of the establishment of citation analysis (Garfield 1955). The six indexes discussed in this paper take into account the two dimensions of the number of publications and citations. To explore the relationship between each index and the number of publications and citations, we used SPSS 21.0 to conduct multiple linear regression analysis. The goodness-of-fit R^2 value and the multiple linear fitting regression coefficients are shown in Table 5.

Table 4 The Pearson correlation coefficients among the six indicators

	h-index	g-index	AR-index	P-index	I3	Academic trace
h-index	1	0.976**	0.942**	0.910**	0.955**	0.829**
g-index	0.976**	1	0.953**	0.903**	0.945**	0.861**
AR-index	0.942**	0.953**	1	0.919**	0.899**	0.859**
p-index	0.910**	0.903**	0.919**	1	0.867**	0.899**
I3	0.955**	0.945**	0.899**	0.867**	1	0.857**
Academic trace	0.829**	0.861**	0.859**	0.899**	0.857**	1

Note that "**" represents a significant correlation at the 0.01 level (2-tailed)

Table 5 The goodness-of-fit R^2 value and multiple linear fitting regression coefficients

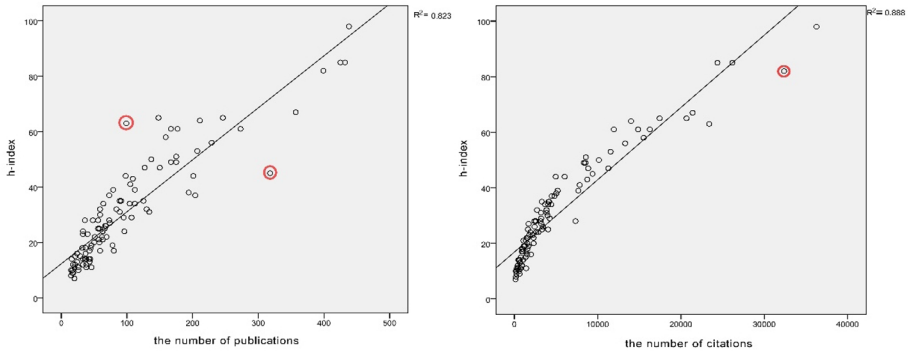
Indicators	R^2	Regression coefficients for the number of publications	Regression coefficients for the number of citations
h-index	0.917	0.073	0.002
g-index	0.884	0.123	0.004
AR-index	0.907	0.018	0.001
p-index	0.860	− 0.127	0.006
I3	0.991	0.724	0.021
academic trace	0.824	− 12.316	0.493

In Table 5, the second column represents the goodness-of-fit R^2 value between each indicator and the number of publications and citations; the third column represents the regression coefficient between each indicator and the number of publications; and the fourth column represents the regression coefficient between each indicator and the number of citations. As shown in Table 5, the goodness-of-fit R^2 value between each indicator and the number of publications and citations is above 0.8, which indicates that the linear fitting effect is good. The absolute value of the regression coefficient between each index and the number of publications is higher than that between each index and the number of citations, which shows that the six indexes are affected by the number of publications more than they are affected by the number of citations. Moreover, the p-index and academic trace negatively correlate with the number of publications, which is easy to understand. According to the p-index formula, if the growth of $N^{1/3}$ exceeds $C^{2/3}$, then the value of the p-index will decrease. The same is true for the academic trace, which covers the whole citation distribution of an author, and if an author has too many zero-cited papers, then the value of the academic trace will decrease. Therefore, if an author blindly pursues having a high number of published papers but the number of citations does not increase, then the value of these two indexes will decline.

Outlier value analysis of the citation distribution

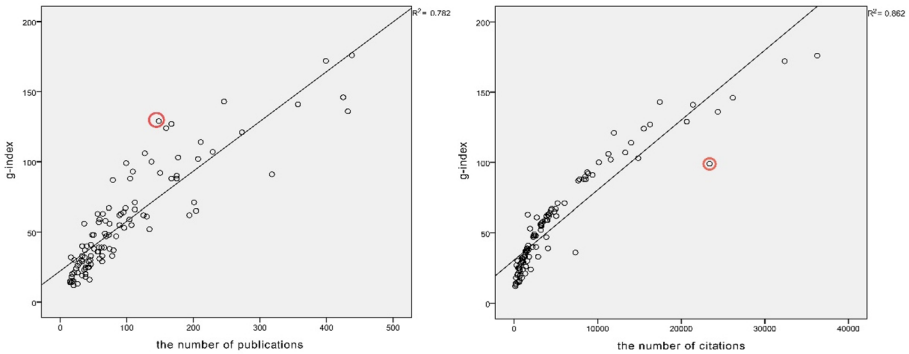
In this section, we draw the scatter plots of the six indicators for the 106 authors. The figure shows that some outliers (marked with red circles) seriously deviate from the trend line. Combined with the citation distribution of the authors' published papers, the limitations of each index in measuring their research performance can be analysed.

In Fig. 2a, the h-index of Farokhzad, Omid C., 63, is higher than that of Aspuru-Guzik, Alan, 45, while the number of papers published by Aspuru-Guzik, Alan, 318, is higher than that by Farokhzad, Omid C., 99. From their citation distribution, Farokhzad, Omid C. published 99 papers, and most of them are highly cited, while Aspuru-Guzik, Alan published 318 papers, with 178 papers being cited no more than ten times and 114 paper receiving zero citations. However, when calculating the h-index of Aspuru-Guzik, Alan, most papers were ignored because each paper was cited no more than 45 times. Moreover, most of these zero-cited or low-cited papers have been published in the past 1 or 2 years. The reason why they did not receive many citations is probably their recent date of publication rather than the poor quality of the papers. The h-index ignores the time factor, which is usually unfair to young scholars. In Fig. 2b, there is an author, Whiteside, George M., whose h-index value and number of citations are 82 and 32,369, respectively. The value of the h-index is



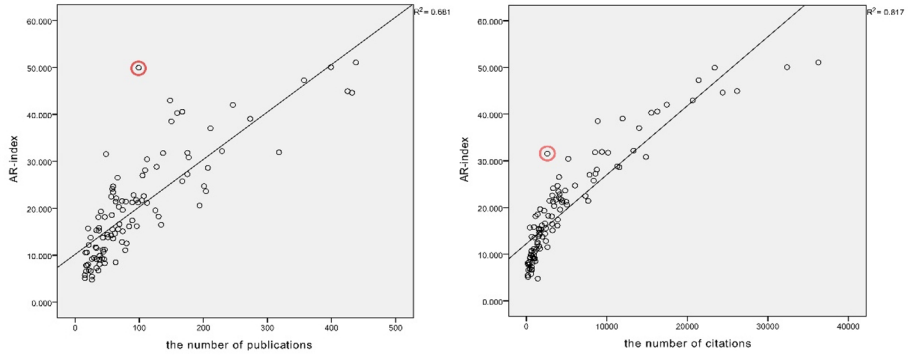
(a) h-index and the number of publications

(b) h-index and the number of citations



(c) g-index and the number of publications

(d) g-index and the number of citations



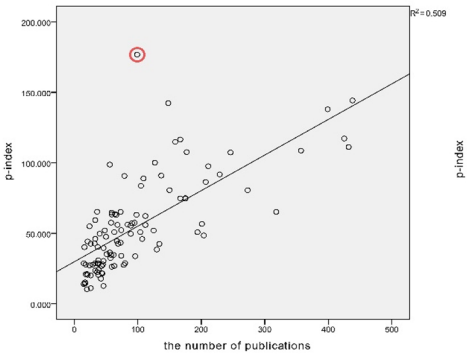
(e) AR-index and the number of publications

(f) AR-index and the number of citations

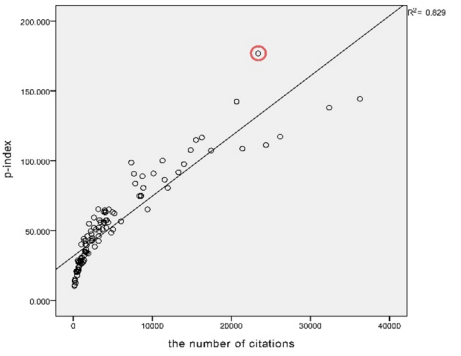
Fig. 2 The scatter plots of the six indicators and the number of publications and citations

less than that of other authors who have the same number of citations. One of his papers received 4098 citations; however, only 82 citations were involved in the calculation of the h-index, and the remaining 4016 citations did not affect the growth of the h-index. Therefore, the h-index is insensitive to highly cited papers (Schreiber 2010a, b; Egghe 2010).

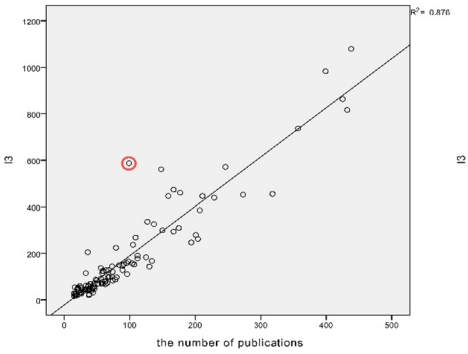
In Fig. 2c, Lieber, Charles M. has a g-index of 143, with a number of papers of 148. His g-index is higher than that of other authors who have published the same number of papers because several of his papers have been cited thousands of times. In Fig. 2d, there is an



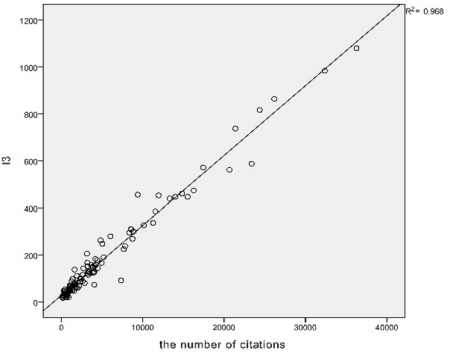
(g) p-index and the number of publications



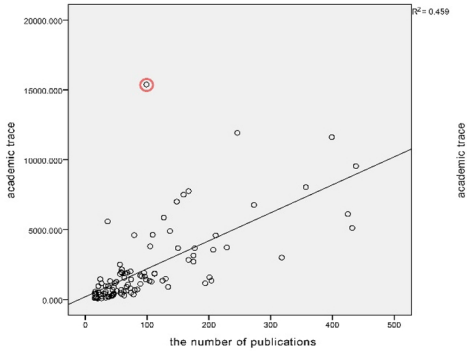
(h) p-index and the number of citations



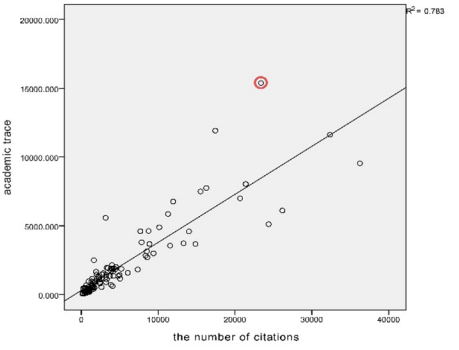
(i) I3 and the number of publications



(j) I3 and the number of citations



(k) academic trace and the number of publications



(l) academic trace and the number of citations

Fig. 2 (continued)

author, Farokhzad, Omid C., whose g-index is 99, which is lower than that of other authors who have the same number of citations because he published only 99 papers. Therefore, if an author publishes a small number of papers but some of his/her papers are highly cited, then the g-index value is often equal to the number of papers. The problem is that once the g-index value equals the number of papers, it will increase not until the author publishes a new paper, even if his/her previous papers receive further citations.

In Fig. 2i, Farokhzad, Omid C., whose AR-index value is 49.946, ranks the third, but his number of papers is only 99, which is far from 438 and 399 papers of the author who ranks the first or second in AR-index value. Because there are 63 papers in the h-core of Farokhzad, Omid C., the 63 papers obtained 22309 citations in total, and the average age of publication is 8.5 years. In addition, there are 36 papers outside of h-core, obtained 876 citations in total. Weissleder, Ralph, ranking the first in AR-index, has 98 papers in h-core, obtaining 25,319 citations in total, with an average publication age of 10 years, and the rest 340 papers have obtained 10,942 citations in total. Therefore, AR-index is worthy of its function of measuring the citation strength in h-core, and it also strengthens the weight of papers published in a short time period. But like h-index, it ignores most of the papers and citations except h-core, which underestimates the influence of many researchers. In Fig. 2j, one author Kats, Mikhail A., whose AR-index is 31.543, is higher than that of the author with the same citation. Although there are only 28 papers in Kats, Mikhail A.'s h-core, 7136 citations have been obtained, with an average publication time of 6.9 years. Another author, Wagner Gerhard, has 44 papers in his h-core, obtaining a total of 4252 citations, with an average publication time of 9.1 years. Therefore, the value of his AR-index is smaller than that of Kats, Mikhail A., which is 24.699. This shows that unlike the h-index, the papers in h-excess still have an effect on the growth of the AR-index, but it will also be affected by the length of publication.

In Fig. 2g, h, there are many outliers, the most typical of which is Farokhzad, Omid C., who has 99 papers and 23,385 citations and whose p-index value is the largest. Hence, the higher the number of citations and the lower the number of papers published, the higher the p-index value an author will obtain. In addition, the p-index focuses only on the number of papers and citations, ignoring the distribution of citations. Thus, as long as the evaluation object has the same number of citations and papers, the p-index value will be the same. To illustrate, suppose that there are author A and author B, who have published 20 papers and received 1000 citations. Author A has received 50 citations per paper; in contrast, author B has published two papers that have been cited 450 times each, with the remaining 18 papers having been cited a total of 100 times. The p-index will treat the two cases equally, failing to highlight the two highly cited papers of author B. Therefore, the p-index is insensitive to highly cited papers.

In Fig. 2i, although Farokhzad, Omid C. published relatively few papers, his I3 score was still large because most of his papers were highly cited. According to the definition of the I3, these papers were divided into categories with high scores; in other words, the I3 is sensitive to highly cited papers. Consequently, the size of the maximum citation in a paper set severely affects the value of the I3. If the maximum citation is too large, then the threshold of classification will be too large. Most papers will be classified as having low scores, as a result of which the I3 value of most authors will be small, thus reducing the degree of discrimination of the I3.

In Fig. 2k, l, Farokhzad, Omid C. has the highest academic trace value. The academic trace considers the entire citation distribution of papers, including the excess of the h-core, h-core and h-tails. Because Farokhzad, Omid C. has a large number of citations in the h-core and a small number of zero-cited papers, he obtains the highest academic trace value in the data set.

Discussion and conclusion

The limitations of the six indicators

With the introduction of an increasing number of scholarly evaluation indicators, researchers are trying to determine their limitations to improve the quality of evaluation. Based on theoretical and empirical analysis, this study compared the evaluation results of the h-index, g-index, AR-index, p-index, I3, and academic trace from three aspects: the degree of discrimination, the consistency of measurement results, the relationship between each index and the number of papers and citations in order to determine their limitations in evaluating authors' research performance. Some of the conclusions of this paper also confirmed previous research results. The h-index value only increases and does not decrease over time, which allows scientists to rest on their laurels (Ye 2014). It is insensitive to highly cited papers and changes in the number of citations (Schreiber 2010a; Egghe 2010), and it ignores papers outside of the h-core, making the measurement result inadequate (Zhang 2013). Once an author's paper is included in the h-core, subsequent citations will not contribute to the growth of the author's h-index. Moreover, the g-index neglects citations of papers outside the g-core (Abramo et al. 2013a, b).

In addition, we found that the six indexes are affected more by the number of publications than by the number of citations. Among the six indexes discussed in this paper, the h-index had the lowest degree of discrimination, followed by the g-index, I3, AR-index, p-index, and academic trace. Compared to the h-index, the accumulation of citation frequency makes it easier for the g-index value to be equal to the number of papers published by an author, and once this equality is reached, subsequent citations received by these papers will no longer contribute to the growth of the g-index unless the author publishes a new paper. The AR-index takes into account the publication age of the paper so that the value of a researcher's AR-index will not only rise blindly, but it does not take into account h-tail's papers and citations, which underestimates the impact of many researchers. The p-index considers only the number of publications and citations, ignoring the citation distribution, which makes the index insensitive to highly cited papers. The measurement result of the I3 is easily affected by changes in the extreme values in a data set, the size of which is the basis for classification. If extreme values are too large, then most papers will be divided into categories with low scores due to an insufficient citation frequency, resulting in a low I3 value for most authors, which also reduces the degree of discrimination of the I3 to an extent. Therefore, the I3 is not suitable for data sets with excessive citation frequency extremums. Additionally, the calculation of the academic trace is relatively complex, which may also affect its application in practice. Table 6 shows the limitations of each indicator in the evaluation.

Of course, these six indicators suffer from the same problems as all simple indicators that use citations. For example, they are field-dependent, may be influenced by self-citations; there is a problem finding reference standards; it is rather difficult to collect all data necessary for the determination of the h-index. Often a scientist's complete publication list is necessary in order to discriminate between scientists with the same name and initial.

Table 6 The limitations of each indicator in evaluation

Indicators	Limitations
h-index	The h-index value only increases and does not decrease over time, which allows scientists to rest on their laurels The h-index ignores most citations and papers outside the h-core, as a result of which the measurement results are not wholly accurate The h-index is insensitive to highly cited papers and changes in the number of their citations The degree of discrimination for the h-index is the lowest among the six indicators
g-index	The g-index value only increases and never declines over time The g-index also ignores citations and papers outside the g-core Compared to the h-index, the accumulation of citation frequency makes it easier for the g-index value to be equal to the number of papers published by an author, and once this equality is reached, subsequent citations received by these papers will no longer contribute to the growth of the g-index unless the author publishes a new paper The degree of discrimination of the g-index is also not good
AR-index	The AR index only considers the impact of h-core and h-excess papers, but it does not consider h-tail papers and citations, which underestimates the impact of many researchers
p-index	The p-index focuses only on the number of citations and papers, ignoring the distribution of citations The p-index is also insensitive to highly cited papers
I3	The I3 is vulnerable to the extremes in a data set, whose measurement result will change with the variation in the maximum citation; thus, it is not suitable for data sets with excessive citation frequency extremums
Academic trace	The calculation of the academic trace is relatively complex, which may also affect its application in practice

Considerations and suggestions regarding the six bibliometric indicators

When evaluating academic performance with these six indicators, the following aspects should be considered. First, there are no perfect indicators, and evaluators should pay attention to their limitations and conditions of applicability when measuring authors' academic performance. Second, it is essential to understand that the measurement results are based on a particular data set; thus, the results may be incomparable to results based on other data sets. Third, to be fair to young scholars, we should pay attention to the cumulative factor of citation frequency over time (Kozak and Bornmann 2012). Fourth, it is not enough to evaluate authors' research performance using only bibliometric indicators. The measurement results should be combined with peer review and even social impact, and a comprehensive academic evaluation should consist of a multi-index, multi-perspective meta-analysis. Finally, the author's influence comes from the diffusion of knowledge in his articles, and the degree of knowledge diffusion should be measured by how many different people, fields, institutions, and regions it is quoted by, and the emphasis is the word "different". However, the existing measures of influence based on citations cannot achieve this goal. Of course, some scholars have also made some attempts in this area, such as Ajiferuke et al. (Ajiferuke and Wolfram 2010; Ajiferuke et al. 2010) created the ch-index based on the citer analysis, and applied it to measure the influence of the author. An article is cited five times by five different author and five times by the same author, the impact is different. Moreover, regardless of how

the methods and indicators of academic evaluation are improved and developed, the purpose of research evaluation is to promote academic progress and scientific development.

Limitations of the study

This study also has some limitations. One hundred and six authors in the field of chemistry from Harvard University were selected as the objects of our empirical analysis; they may constitute a convenient sample, but such a sample may also limit the universality of our research results. However, the size of data samples is not the most important factor in revealing the limitations of the six indicators. Nonetheless, in the future, we will choose different source data to validate our findings.

Acknowledgements We wish to thank the anonymous referees for important insightful comments and suggestions. This research was funded by Project of the National Social Science Foundation of China (Grant No. 17BTQ071).

Author contribution JD was involved in conceptualization, formal analysis, methodology, supervision, review, the preparation of the initial draft, and the editing of the final draft. CL was involved in data collection, formal analysis, methodology, software, and the preparation of the initial draft. GK was involved in formal analysis, review and the editing of the final draft.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

Abramo, G., D'Angelo, C. A., & Viel, F. (2013a). Assessing the accuracy of the h- and g-indexes for measuring researchers' productivity. *Journal of the American Society for Information Science and Technology, 64*(6), 1224–1234.

Abramo, G., D'Angelo, C. A., & Viel, F. (2013b). The suitability of h and g indexes for measuring the research performance of institutions. *Scientometrics, 97*(3), 555–570.

Agarwal, A., Durairajanayagam, D., Tatagari, S., et al. (2016). Bibliometrics -tracking research impact by selecting the appropriate metrics. *Asian Journal of Andrology, 18*(2), 296–309.

Ajiferuke, I., Lu, K., & Wolfram, D. (2010). A comparison of citer and citation-based measure outcomes for multiple disciplines. *Scientometrics, 61*(10), 2086–2096.

Ajiferuke, I., & Wolfram, D. (2010). Citer analysis as a measure of research impact: Library and information science as a case study. *Scientometrics, 83*(3), 623–638.

Alonso, S., Cabrerizo, F. J., Herrera-Viedma, E., et al. (2010). Hg-index: A new index to characterize the scientific output of researchers based on the h-and g-indices. *Scientometrics, 82*(2), 391–400.

Ball, P. (2005). Index aims for fair ranking of scientists. *Nature, 436*(7053), 900.

Barnes, C. (2017). The h-index debate: An introduction for librarians. *Journal of Academic Librarianship, 43*(6), 487–494.

Batista, P. D., Campiteli, M. G., & Kinouchi, O. (2006). Is it possible to compare researchers with different scientific interests? *Scientometrics, 68*(1), 179–189.

Bornmann, L. (2013). How to analyze percentile citation impact data meaningfully in bibliometrics: The statistical analysis of distributions, percentile rank classes, and top-cited papers. *Journal of the American Society for Information Science and Technology, 64*(3), 587–595.

Bornmann, L. (2014). H-Index research in scientometrics: A summary. *Journal of Informetrics, 3*(8), 749–750.

Bornmann, L., Mutz, R., Hug, S., et al. (2011). A multilevel meta-analysis of studies reporting correlations between the h index and 37 different h index variants. *Journal of Informetrics, 5*(3), 346–359.

Bornmann, L., Rüdiger, M., & Daniel, H. D. (2008). Are there better indices for evaluation purposes than the h index? A comparison of nine different variants of the h index using data from biomedicine. *Journal of the American Society for Information Science and Technology, 59*(5), 830–837.
Burrell, Q. L. (2007). On the h-index, the size of the Hirsch core and Jin's A-index. *Journal of Informetrics, 1*(2), 170–177.
Costas, R., & Bordons, M. (2008). Is g-index better than h-index? An exploratory study at the individual level. *Scientometrics, 77*(2), 267–288.
Egghe, L. (2006). Theory and practice of the g-index. *Scientometrics, 69*(1), 131–152.
Egghe, L. (2010). The Hirsch index and related impact measures. *Annual Review of Information Science and Technology, 44*(1), 65–114.
Garfield, E. (1955). Citation indexes for science. *Science, 122*(3159), 108–111.
Gingras, Y., & Larivière, V. (2011). There are neither "king" nor "crown" in scientometrics: Comments on a supposed "alternative" method of normalization. *Journal of Informetrics, 5*(1), 226–227.
Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceeding of the National Academy of Sciences of USA, 102*(46), 16569–16572.
Hirsch, J. E. (2010). An index to quantify an individual's scientific research output that takes into account the effect of multiple coauthorship. *Scientometrics, 85*(3), 741–754.
Hirsch, J. E. (2019). hα: An index to quantify an individual's scientific leadership. *Scientometrics, 118*(2), 673–686.
Iglesias, J. E., & Pecharromán, C. (2007). Scaling the h-index for different scientific ISI fields. *Scientometrics, 73*(3), 303–320.
Jin, B. H. (2006). H-index: An evaluation indicator proposed by scientist. *Science Focus (in Chinese), 1*(1), 8–9.
Jin, B. H., Liang, L. M., Rousseau, R., et al. (2007). The R-and AR indices: Complementing the h-index. *Chinese Science Bulletin, 52*(6), 855–863.
Kozak, M., & Bornmann, L. (2012). A new family of cumulative indexes for measuring scientific performance. *PLoS ONE, 7*(10), e47679.
Leydesdorff, L., & Bornmann, L. (2011). Integrated impact indicators compared with impact factors: An alternative research design with policy implications. *Journal of the American Society for Information Science and Technology, 62*(11), 2133–2146.
Liang, L. (2006). h-index sequence and h-index matrix: Constructions and applications. *Scientometrics, 69*(1), 153–159.
Prathap, G. (2010). Is there a place for a mock h-index? *Scientometrics, 84*(1), 153–165.
Rousseau, R. (2006). Simple models and the corresponding h- and g-index. Retrieved March 11, 2019, from http://eprints.rclis.org/7501/1/Rousseau_Dalian.pdf.
Rousseau, R., & Ye, F. Y. (2012). A formal relation between the h-index of a set of articles and their I3 score. *Journal of Informetrics, 6*(1), 34–35.
Schreiber, M. (2008a). An empirical investigation of the g-index for 26 physicists in comparison with the h-index, the A-index, and the R-index. *Journal of the American Society for Information Science and Technology, 59*(9), 1513–1522.
Schreiber, M. (2008b). To share the fame in a fair way, hm modifies h for multi-authored manuscripts. *New Journal of Physics, 10*(4), 1–9.
Schreiber, M. (2009). Fractionalized counting of publications for the g-Index. *Journal of the American Society for Information Science and Technology, 60*(10), 2145–2150.
Schreiber, M. (2010a). Twenty Hirsch index variants and other indicators giving more or less preference to highly cited papers. *Annalen der Physik, 522*(8), 536–554.
Schreiber, M. (2010b). Revisiting the g-index: The average number of citations in the g-core. *Journal of the American Society for Information Science and Technology, 61*(1), 169–174.
Sidiropoulos, A., Katsaros, D., & Manolopoulos, Y. (2007). Generalized Hirsch h-index for disclosing latent facts in citation networks. *Scientometrics, 72*(2), 253–280.
Tol, R. S. J. (2008). A rational, successive g-index applied to economics departments in Ireland. *Journal of Informetrics, 2*(2), 149–155.
Wagner, C. S., & Leydesdorff, L. (2012). An Integrated Impact Indicator: A new definition of 'Impact' with policy relevance. *Research Evaluation, 21*(3), 183–188.
Woeginger, G. J. (2009). Generalizations of Egghe's g-index. *Journal of the American Society for Information Science and Technology, 60*(6), 1267–1273.
Ye, F. Y. (2014). Overview of research status and development of international academic evaluation indicators. *Journal of Intelligence, 33*(2), 215–223.

Ye, F. Y., Bornmann, L., & Leydesdorff, L. (2017). h-based I3-type multivariate vectors: Multidimensional indicators of publication and citation scores. *COLLNET Journal of Scientometrics and Information Management, 11*(1), 153–171.

Ye, F. Y., & Leydesdorff, L. (2014). The "Academic Trace" of the performance matrix: A mathematical synthesis of the h-index and the integrated impact indicator (I3). *Journal of the Association for Information Science and Technology, 65*(4), 742–750.

Zhang, C. T. (2009a). The e-index, complementing the h-index for excess citations. *PLoS ONE, 4*(5), e5429.

Zhang, C. T. (2009b). A proposal for calculating weighted citations based on author rank. *EMBO Reports, 10*(5), 416–417.

Zhang, C. T. (2013). The h'-index, effectively improving the h-index based on the citation distribution. *PLoS ONE, 8*(4), e59912.