Comparative study of four impact measures and qualitative conclusions

by

L. Egghe

Universiteit Hasselt (UHasselt), Campus Diepenbeek, Agoralaan, B-3590 Diepenbeek, Belgium[1]
and
Universiteit Antwerpen (UA), IBW, Stadscampus, Venusstraat 35, B-2000 Antwerpen, Belgium

leo.egghe@uhasselt.be

ABSTRACT

We present a comparative study of four impact measures: the *h*-index, the *g*-index, the *R*-index and the *j*-index. The *g*-index satisfies the transfer principle, the *j*-index satisfies the opposite transfer principle while the *h*- and *R*-indices do not satisfy any of these principles. We study general inequalities between these measures and also determine their maximal and minimal values, given a fixed total number of citations

---

[1] Permanent address

Key words and phrases: *h*-index, Hirsch, *g*-index, *R*-index, *j*-index, transfer principle

Introduction

Let us have a set of *n* papers each having a certain number of citations, denoted c_i, $i = 1,...,n$. We suppose that these papers are ranked in decreasing order of their number of citations: $i < j$ implies $c_i \geq c_j$. This set of papers can be the collection of papers of an author or of a journal in a certain period. In fact this framework can be generalized into a set of *n* sources each having a certain number of items, denoted c_i, $i = 1,...,n$ (Egghe (2005, 2010)) but in this paper we will restrict our terminology to the papers-citations relation.

In this framework one can define so-called impact measures. It all started with the definition of the famous Hirsch-index (or *h*-index) which is defined as the highest rank $r = h$ such that all papers on ranks $1,...,h$ each have at least *h* citations. The literature on this index is vast: we refer to Egghe (2010) for a review. We do not go into the pros and cons of the *h*-index, except for one disadvantage (since we need it to understand the other three impact measures): the *h*-index is not sensitive to the actual number of citations for the papers in the *h*-core (the papers on ranks $1,...,h$): once a paper is in the *h*-core (hence for which $c_i \geq h$ if the paper has rank *i*), it does not matter how large c_i is. This means that highly cited papers are not counted as such in the calculation of the *h*-index and this is a well-known disadvantage of this index.

For this reason, Egghe (2006) introduced the *g*-index. In the same ranking of the papers as above, the *g*-index is the highest rank $r = g$ such that

$$\sum_{i=1}^{r} g_i \geq r^2 \qquad (1)$$

This was inspired by the fact that the *h*-index satisfies this inequality (and hence $h \leq g$, by definition of the *g*-index) and the fact that we now, effectively, use the actual number of citations in the highly cited papers. This definition is equivalent with the following: the *g*-index is the highest rank $r = g$ such that the average number of citations to the first *r* papers is at least *r* (note that in the definition of the *h*-index, this is required for any paper in the first

r ranks). It is explained in Egghe (2006) that the *g*-index is indeed sensitive to the number of citations to the highest cited papers.

The *R*-index, introduced in Jin et al. (2007), has the same goal as the *g*-index: solving the above discussed disadvantage of the *h*-index. Its definition uses the *h*-index itself and reads as in equation (2).

$$R = \sqrt{\sum_{i=1}^{h} c_i} \tag{2}$$

So, it is the square root of the total number of citations to the papers in the *h*-core. As the *g*-index it is indeed sensitive to the number of citations to the highest cited papers.

A very recently introduced impact measure, striving after the same goals as the *g*- and *R*-index, is the *j*-index (Levene, Fenner and Bar-Ilan (2012)). Its definition is given in formula (3)

$$j = \sum_{i=1}^{n} \sqrt{c_i} \tag{3}$$

Hence it is the sum of the square roots of the number of citations to each paper. Again one can see that the *j*-index is sensitive to the number of citations to the highest cited papers. The calculation of (3) requires, however, that c_i is known for all papers.

The next section is devoted to the study of the transfer property, well-known from econometrics (Egghe (2005, 2009)). In econometric terms it states that, in a community, if one takes money away from a poor person and gives it to a richer person, the inequality in this community has increased. In our framework, where c_i is the number of citations to the i^{th} paper and where one ranks papers in decreasing order of this number this means that one studies the following transformation. Let $(c_1,...,c_n)$ be the original decreasing vector as above (so $c_i \geq c_l$ for $i < l$). We then take such a couple (i,l) and to c_i one adds a positive number *k*

and from c_l one subtracts this number k (such that $k \le c_l$ of course) in order to become a vector

$$(c_1,....,c_i+k,....,c_l-k,....,c_n) \tag{4}$$

This is called an elementary transfer.

What does such a transfer mean in terms of citations? Remark that in (4) one has the same number of papers as in $(c_1,...,c_i,...,c_l,...,c_n)$ as well as the same total number of citations. However, in (4), the citations are more concentrated over the papers. We can say that (4) represents a more "elitary" situation. A repeated application of transfers as in (4) yields a publication – citation situation where some papers attract lots of citations and some papers are lowly cited. We could say that in such a situation some papers are very important, i.e. more important than in the starting situation $(c_1,...,c_i,...,c_l,...,c_n)$.

In the next section we will study the *h*-, *g*-, *R*- and *j*-indices in this context. Do these indices increase or decrease when applying such a transformation (4) or not? An answer will be given in the next section.

In the third section we will derive from these results, the maximum and minimum values of these indices, given that

$$C=\sum_{i=1}^{n} c_i \tag{5}$$

, the total number of citations, is constant (note that in case (4) the total number of citations remained constant). Further in the third section we will prove general inequalities between the *h*-, *g*-, *R*- and *j*-indices.

Then follows the conclusions section with some advises for further research.

The transfer property

In econometrics, the transfer property refers to the fact that a transfer as in (4) increases the inequality (concentration) of the system and, hence, any good concentration measure *f* should increase:

$$f\left(c_1,...,c_i+k,...,c_l-k,...,c_n\right)\geq f\left(c_1,...,c_i,...,c_l,...,c_n\right) \tag{6}$$

If the opposite happens (see (7)) we say that *f* is a good measure of diversity, a desired property in biometrics (Rousseau and Van Hecke (1999)):

$$f\left(c_1,...,c_i+k,...,c_l-k,...,c_n\right)\leq f\left(c_1,...,c_i,...,c_l,...,c_n\right) \tag{7}$$

The following proposition on the *g*-index appeared already in Egghe (2009).

Proposition 1 (Egghe (2009)): The *g*-index is a good measure of concentration.

Proof: The proof is very short. Since in (4), denoting this vector as $\left(c'_1,...,c'_n\right)$, no partial sum

$$\sum_{l=1}^{m} c'_l$$

is smaller than the corresponding partial sum

$$\sum_{l=1}^{m} c_l$$

of the original vector, the definition of the *g*-index shows that (6) is valid with *f* replaced by the *g*-index. □

For the *j*-index we have the following opposite result.

Proposition 2: The *j*-index is a good measure of diversity and where the inequality in (7) is even strict:

$$j\left(c_1,...,c_i+k,...,c_l-k,...,c_n\right)<j\left(c_1,...,c_i,...,c_l,...,c_n\right) \tag{8}$$

Proof: We have to show, by (3), that

$$\sum_{\substack{m=1\\m\neq i,l}}^{n}\sqrt{c_m}+\sqrt{c_i+k}+\sqrt{c_l-k}<\sum_{m=1}^{n}\sqrt{c_m}$$

Hence we have to show

$$\sqrt{c_i+k}+\sqrt{c_l-k}<\sqrt{c_i}+\sqrt{c_l} \tag{9}$$

The square of the left hand side of (9) is equal to

$$c_i+c_l+2\sqrt{\left(c_i+k\right)\left(c_l-k\right)}$$

while the square of the right hand side of (9) is equal to

$$c_i+c_l+2\sqrt{c_ic_l}$$

Hence we have to show that

$$\sqrt{\left(c_i+k\right)\left(c_l-k\right)}<\sqrt{c_ic_l} \tag{10}$$

But

$$\begin{aligned}&(c_i + k)(c_l - k)\\&= c_i c_l - kc_i + kc_l - k^2\\&= c_i c_l - k(c_i - c_l) - k^2\\&< c_i c_l\end{aligned}$$

since $c_i \geq c_l$, since we have the decreasing vector $(c_1, ..., c_n)$. □

The *h*-index and *R*-index do not satisfy these properties as the next examples show.

Example 3: The *h*-index and *R*-index are not good concentration measures.

Indeed, let $X = (3,3,3)$ the citation vector. Hence its *h*-index is $h(X) = 3$ and its *R*-index is $R(X) = \sqrt{9} = 3$. Then apply the elementary transfer of one unit from the third article to the first one. This yields $Y = (4,3,2)$ and hence $h(Y) = 2 < h(X)$ and $R(Y) = \sqrt{7} < R(X)$ contradicting the properties of good concentration measures.

One might now think that the *h*-index and *R*-index are good diversity measures but also that is not true as the next example shows.

Example 4: The *h*-index and *R*-index are not good diversity measures.

Indeed, let $X = (3,3,2,1)$. Hence $h(X) = 2$ and $R(X) = \sqrt{6}$. Then apply the elementary transfer of one unit from the fourth article to the third one, yielding $Y = (3,3,3,0)$. Now $h(Y) = 3 > h(X)$ and $R(Y) = \sqrt{9} > R(X)$.

This clearly shows the different properties of these four measures. We will not go into the debate of which properties are the best but one should be aware that the *g*-index favors concentration of citations over articles (advocated in Egghe (2009)) while the *j*-index favors a more equal spread out of citations over articles, a property defended in De Visscher (2011) but see also Egghe (2012).

Remark: If we take the measures $C = \sum_{i=1}^{n} c_i$ or $\sqrt{C}$ it is clear that they are invariant under an elementary transfer.

Based on the above results we will, in the next section, determine maximum and minimum values of these indices, which will again show the different nature of these indices.

Relations between the h-index, g-index, R-index and j-index

We have the following corollary of Proposition 1.

Corollary 5: Given a fixed total number *C* of citations (as in (5)), we have, if $n \geq \sqrt{C}$

(i) max $g = \left[\sqrt{C}\right]$ ($[x]$ denotes the largest entire number smaller than or equal to *x*) and is yielded for the vector $(c_1,...,c_n)$ where $c_1 = C,\ c_2 = ... = c_n = 0$.

(ii) min $g = \left[\frac{C}{n}\right]$ and is yielded for the vector $(c_1,...,c_n)$ where $c_1 = ... = c_n = \frac{C}{n}$.

Proof:

By Proposition 1, the g-index increases with elementary transfers. Hence the *g*-index is maximal for the citation vector $(C,0,...,0)$ and is minimal for the citation vector $\left(\frac{C}{n},...,\frac{C}{n}\right)$.

(i) For the vector $(C,0,...,0)$ we have, by definition of the *g*-index, that *g* is the largest rank such that $C \geq g^2$. Hence *g* is the largest rank such that $g \leq \sqrt{C}$. But since *g* is an entire number and since $n \geq \sqrt{C}$ we have that $g = \left[\sqrt{C}\right]$. Hence max $g = \left[\sqrt{C}\right]$.

(ii) For the vector $\left(\frac{C}{n},...,\frac{C}{n}\right)$ we have that the *g*-index is the largest rank such that $\frac{C}{n} g \geq g^2$ hence $g \leq \frac{C}{n}$. Since *g* is an entire number and since $n \geq \sqrt{C}$ we have $g = \left[\frac{C}{n}\right]$. Hence min $g = \left[\frac{C}{n}\right]$. □

For the *j*-index we have the opposite result.

Corollary 6: Given a fixed total number *C* of citations (as in (5)), we have

(i) max $j = \sqrt{nC}$ and is yielded for the vector $(c_1,...,c_n)$ where $c_1 = ... = c_n = \frac{C}{n}$.

(ii) min $j = \sqrt{C}$ and is yielded for the vector $(c_1,...,c_n)$ where $c_1 = C$, $c_2 = ... = c_n = 0$.

Proof:

By Proposition 2, the *j*-index decreases with elementary transfers. Hence the *j*-index is maximal for the citation vector $\left(\frac{C}{n},...,\frac{C}{n}\right)$ and is minimal for the citation vector $(C,0,...,0)$.

(i) For the vector $\left(\frac{C}{n},...,\frac{C}{n}\right)$ we have, by definition of the *j*-index, that $j = n\sqrt{\frac{C}{n}} = \sqrt{nC}$. Hence max $j = \sqrt{nC}$.

(ii) For the citation vector $(C,0,...,0)$ we, have, by definition of the *j*-index, that $j=\sqrt{C}$. Hence min $j=\sqrt{C}$. □

As is clear from the two corollaries above, the *g*-index and *j*-index are very different impact measures, what was already clear from Proposition 1 and 2. For the *h*- and *R*-index we cannot use these propositions since they do not satisfy them. We have the following results.

Proposition 7: Given a fixed total number *C* of citations (as in (5)), we have

(i) max $h=\left[\sqrt{C}\right]$ and is yielded for the vector $(c_1,...,c_n)$ where

$$c_1=...=c_{\left[\sqrt{C}\right]}=\left[\sqrt{C}\right] \tag{11}$$

(ii) min $h=1$ and is yielded for the vector $(c_1,...,c_n)$ where $c_1=C$ and $c_2=...=c_n=0$.

Proof:

(i) *h* can never be larger than $\sqrt{C}$, by definition of *h*. Hence $h\le\sqrt{C}$. But since *h* is an entire number, $h\le\left[\sqrt{C}\right]$. We can, actually reach $h=\left[\sqrt{C}\right]$ by the vector in (11) (since we do not use more than *C* citations). Hence max $h=\left[\sqrt{C}\right]$.

(ii) Is trivial □

Proposition 8: Given a fixed total number *C* of citations (as in (5)), we have

(i) max $R=\sqrt{C}$ which is yielded for the vector $(C,0,...,0)$.

(ii) min $R=\sqrt{h\frac{C}{n}}$ and is yielded for the vector $(c_1,...,c_n)$ where $c_1=...=c_n=\frac{C}{n}$.

Proof:

(i) Is trivial.

(ii) R is minimal if there is a maximum of citations on the ranks $h+1,...,n$. Since the vector $(c_1,...,c_n)$ decreases, this is reached if the vector is constant: $c_1 = ... = c_n = \frac{C}{n}$. By definition of the R-index we now have $R = \sqrt{h\frac{C}{n}}$. Hence $\min\ R = \sqrt{h\frac{C}{n}}$.

Note that, since $h \le n$, $\min R \le \max R$, as it should.

□

Note that also the h- and R-index are different from the other indices, although the properties of the R-index are somewhat similar to these of the g-index.

We close this paper by proving some general inequalities between these indices.

Proposition 9: In general we have the following inequalities between the h-index, g-index, R-index and j-index

$$h \le \lceil R \rceil \le R \le \sqrt{C} \le j \tag{12}$$

$$h \le g \le \left\lfloor \sqrt{C} \right\rfloor \le \sqrt{C} \le j \tag{13}$$

Proof:

Since all $c_i \ge h$ for $i = 1,...,h$ we have that $h \le R$. Since h is an entire number we hence have $h \le \lceil R \rceil$. That $R \le \sqrt{C}$ is trivial. That $\sqrt{C} \le j$ is trivial from the definition of the j-index.

That $g \le \left\lfloor \sqrt{C} \right\rfloor$ follows from Corollary 5. □

Remark:

Since

$$R = \sqrt{\sum_{i=1}^{h} c_i}$$

and since g is the largest rank such that

$$\sum_{i=1}^{g} c_i \geq g^2$$

hence

$$\sqrt{\sum_{i=1}^{g} c_i} \geq g$$

hence (since g is an entire number)

$$\left[\sqrt{\sum_{i=1}^{g} c_i}\right] \geq g$$

and since $g \geq h$, one would be inclined to think that $[R] \leq g$.

Here is a counterexample: take the citation vector (9,3). Hence $h = 2$, $g = 2$, $R = \sqrt{12}$, $[R] = 3 > g$.

Conclusions

We have (re-)defined the transfer property. If, with such an elementary transfer, a measure increases we say that it is a good measure of concentration (or inequality). The *g*-index is such a good measure of concentration. If, with such an elementary transfer, a measure decreases we say that it is a good measure of diversity. The *j*-index is such a good measure of diversity. We also show by example that neither the *h*-index nor the *R*-index satisfy these two properties.

Although the *g*-, *R*- and *j*-indices were defined to be more sensitive to the number of citations to the highest cited papers than the *h*-index, the above shows that these measures are very different. We also calculate the maximum and minimum values of all four indices from which again follow their differences.

Finally we prove some general inequalities between these four indices.

We leave open to conduct similar studies of other impact measures.

References

A. De Visscher (2011). What does the *g*-index really measure? Journal of the American Society for Information Science and Technology 62 (11), 2290-2293.

L. Egghe (2005). Power Laws in the Information Production Process: Lotkaian Informetrics. Elsevier, Oxford, UK.

L. Egghe (2006). Theory and practise of the *g*-index. Scientometrics 69(1), 131-152.

L. Egghe (2009). An econometric property of the *g*-index. Information Processing and Management 45(4), 484-489.

L. Egghe (2010). The Hirsch-index and related impact measures. Annual Review of Information Science and Technology, Volume 44 (B. Cronin, ed.), 65-114, Information Today, Inc., Medford, New Jersey, USA.

L. Egghe (2012). Remarks on a paper of A. De Visscher "What does the *g*-index really measure?". Journal of the American Society for Information Science and Technology, 63(10), 2118-2121, 2012.

B. Jin, L. Liang, R. Rousseau and L. Egghe (2007). The *R*- and *AR*-indices: Complementing the *h*-index. Chinese Science Bulletin 52(6), 855-863.

M. Levence, T. Fenner and J. Bar-Ilan (2012). A bibliometric index based on the complete list of cited publications. Cybermetrics 16, Issue 1, Paper 1.

R. Rousseau and P. Van Hecke (1999). Measuring biodiversity. Acta Biotheoretica 47, 1-5.