

Real and rational variants of the h -index and the g -index

Raf Guns¹ and Ronald Rousseau^{2,3,4}

¹ *E-mail: raf.guns@ua.ac.be*

University of Antwerp, IBW, Venusstraat 35, City Campus, 2000 Antwerpen, Belgium

² *E-mail: ronald.rousseau@khbo.be*

KHBO (Association K.U.Leuven), Industrial Sciences and Technology, Zeedijk 101,
B-8400 Oostende, Belgium

³ Hasselt University, Universitaire Campus, B-3590 Diepenbeek, Belgium

⁴ K.U.Leuven, Steunpunt O&O Indicatoren and Dept. MSI,
Dekenstraat 2, B-3000 Leuven, Belgium

Abstract

The definitions of the rational and real-valued variants of the h -index and g -index are reviewed. It is shown how they can be obtained both graphically and by calculation. Formulae are derived expressing the exact relations between the h -variants and between the g -variants. Subsequently these relations are examined. In a citation context the real h -index is often, but not always, smaller than the rational h -index. It is also shown that the relation between the real and the rational g -index depends on the number of citations of the article ranked $g+1$. Maximum differences between h , h_r and h_{rat} on the one hand and between g , g_r and g_{rat} on the other are determined.

1. Introduction

If a scientist's publications are ranked in decreasing order of number of citations, then this person's lifetime achievement h -index is the highest rank such that the first h publications each received h or more citations (Hirsch, 2005). One of the advantages of the h -index is its robustness, in that it is insensitive to low-impact publications with few or no citations.

On the other hand the h -index is also insensitive to exceptional publications: as soon as such a publication is part of the h -core (the group of the h most highly cited publications), its actual number of citations has no longer an influence. Egghe (2006a, b) introduced the g -index to overcome this potential disadvantage. The g -index is the highest rank such that the first g publications together have at least g^2 citations. If there are N publications in total and N^2 is less than the sum of all citations, one adds fictitious articles with zero citations in order to determine the g -index.

Let $P(r)$ denote the number of citations of the r th publication (in general terms: the production of the r th source). For the sake of readability we will denote the cumulative number of citations of the r th publication as $Q(r) = \sum_{i=1}^r P(i)$. The h -index is characterized by the inequalities

$$h \leq P(h) \text{ and } P(h+1) < h + 1 \quad (1)$$

The g -index is characterized by the following inequality.

$$g^2 \leq Q(g) \leq Q(g+1) < (g+1)^2 \quad (2)$$

Consider the simple example in Table 1. It can readily be seen that $h = 4$ and $g = 5$. The publication at rank 6 could be a fictitious publication or a real publication with zero citations; for the calculation of the g -index and related indices, this is completely equivalent.

In this article we present a closer look at the rational and real-valued variants of the h -index and g -index. These variants are (different types of) interpolations between h and $h+1$ and between g and $g+1$ respectively. The rational variants are mainly a refinement of h and g , indicating how close one is to achieving a higher h - or g -index. The real variants are especially useful when one wants to calculate h or g for data that are not natural numbers, for instance because one uses a correction factor to account for collaboration (Chai et al., 2008).

In sections 2 and 3, we illustrate how these indices can be calculated and how they can be represented graphically. This leads to an examination of their mathematical relations in section 4. Since the relations are relatively complex, there is no universally true relation, but it can be shown that in a publication–citation context the real variants are typically smaller than the rational ones. Finally, in section 5 the maximum differences between the variants of the h -index and between the variants of the g -index are determined. It is shown that these maximum differences all approach one, as h or g grow larger.

Table 1. A simple example illustrating the h -index and g -index

r	$P(r)$	r^2	$Q(r)$
1	9	1	9
2	7	4	16
3	6	9	22
4	5	16	27
5	3	25	30
6	0	36	30

2. The rational h -index and g -index

The rational variants of the h -index, denoted as h_{rat} , and the g -index, denoted as g_{rat} , were introduced by Ruane and Tol (Ruane & Tol, 2008; Tol, 2008). As already mentioned, these variants interpolate between h and $h+1$ and between g and $g+1$ respectively. Intuitively speaking, they indicate the 'distance' to a higher h - or g -index. They are only defined in case all $P(r)$ are natural numbers.

Definition of h_{rat} : Consider a researcher with h -index h . Let n be the smallest number of citations that this researcher needs to reach an h -index equal to $h+1$, and let $n_{max} = 2h + 1$, then h_{rat} is defined as:

$$h_{rat} = h + 1 - \frac{n}{n_{max}} = h + 1 - \frac{n}{2h + 1} \quad (3)$$

If a researcher has h -index h , then the value $2h + 1$ is the largest possible minimum citation increment necessary to reach an h -index equal to $h + 1$. This happens in the 'worst case scenario' that there are h publications with h citations each and the publication at rank $h + 1$ has 0 citations. In this 'worst case', the first h publications each need one extra citation while the publication at rank $h + 1$ needs $h + 1$ extra citations.

For a given g , the 'worst case scenario' is

$$g^2 = Q(g) = Q(g+1) < (g+1)^2 = g^2 + 2g + 1$$

and, thus, the maximum number of citations needed to obtain a higher g -index is equal to $2g + 1$. Consequently, we obtain the following definition.

Definition of g_{rat} : Given a researcher with g -index g , then the rational g -index g_{rat} is defined as:

$$g_{rat} = g + 1 - \frac{n}{n_{max}} = g + 1 - \frac{(g+1)^2 - Q(g+1)}{2g+1} = g + \frac{Q(g+1) - g^2}{2g+1} \quad (4)$$

Applying these definitions to the example of Table 1, we get the following results:

$$h_{rat} = 4 + 1 - \frac{2}{8+1} = 4.78$$

$$g_{rat} = 5 + 1 - \frac{36-30}{10+1} = 5.45$$

3. The real-valued h -index and g -index

The real-valued (in short: real) h -index h_r and g -index g_r were introduced by Rousseau (2006) as a generalization of the original indices. The real variants can also be used when citation scores are not natural numbers, for instance when citations are counted fractionally. The real h -index h_r is used in the definition of the adapted pure h -index, a proposal to take co-authors into account when calculating a scientist's h -index (Chai et al., 2008).

Definition of h_r : Let $P(r)$ denote the number of citations of the r th publication and let $P(x)$ denote its piecewise linear interpolation, then the real h -index h_r is the abscissa of the intersection of the function $P(x)$ and the angle bisector $y = x$.

Definition of g_r : Let $Q(r)$ denote the cumulative citation count of all publications up to (and including) r , i.e. $Q(r) = \sum_{i=1}^r P(i)$, and let $Q(x)$ denote its piecewise linear interpolation, then the real g -index g_r is the abscissa of the intersection of the function $Q(x)$ and the curve $y = x^2$.

Given the example in Table 1, h_r and g_r can be determined graphically, as shown in Figure 1. They can also be calculated, as we will now illustrate.

We first determine the formula for h_r . As $h \leq h_r$ and $P(h+1) < h + 1$, this intersection is situated on the line segment connecting $(h, P(h))$ and $(h + 1, P(h+1))$. Consequently: $h \leq h_r < h + 1$.

The line connecting $(h, P(h))$ and $(h + 1, P(h+1))$ has the following equation:

$$y = P(h) + \frac{P(h+1) - P(h)}{(h+1) - h} (x - h)$$

or:

$$y = x \cdot (P(h+1) - P(h)) + (h+1) \cdot P(h) - h \cdot P(h+1) \quad (5)$$

The abscissa of the intersection of this line and the angle bisector $y = x$ is:

$$x = x \cdot (P(h+1) - P(h)) + (h+1) \cdot P(h) - h \cdot P(h+1)$$

$$\Leftrightarrow x = \frac{(h+1) \cdot P(h) - h \cdot P(h+1)}{1 - P(h+1) + P(h)}$$

and hence:

$$h_r = \frac{(h+1) \cdot P(h) - h \cdot P(h+1)}{1 - P(h+1) + P(h)} \quad (6)$$

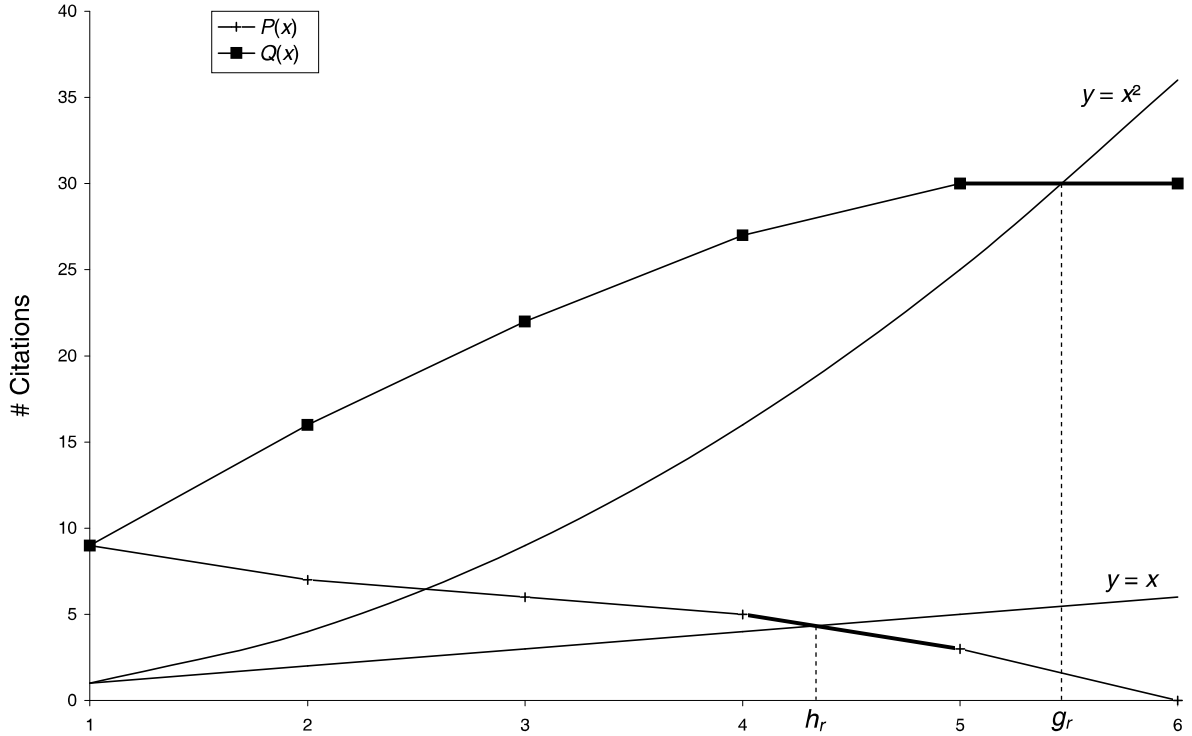


Figure 1. Graphical construction for the calculation of h_r and g_r using the example of Table 1 (the horizontal axis is the x-axis, the vertical axis is the y-axis)

The formula for g_r can be obtained in a similar way. The intersection is located on the line segment that connects $(g, Q(g))$ and $(g + 1, Q(g+1))$. The line connecting these points has the equation:

$$y = x \cdot (Q(g+1) - Q(g)) + (g+1) \cdot Q(g) - g \cdot Q(g+1) \quad (7)$$

The intersection of this line with $y = x^2$ is:

$$x^2 = x \cdot (Q(g+1) - Q(g)) + (g+1) \cdot Q(g) - g \cdot Q(g+1)$$

$$\Leftrightarrow x = \frac{Q(g+1) - Q(g) + \sqrt{4[(g+1) \cdot Q(g) - g \cdot Q(g+1)] + (Q(g) - Q(g+1))^2}}{2}$$

and hence:

$$g_r = \frac{Q(g+1) - Q(g) + \sqrt{4[(g+1) \cdot Q(g) - g \cdot Q(g+1)] + (Q(g) - Q(g+1))^2}}{2} \quad (8)$$

or:

$$g_r = \frac{P(g+1) + \sqrt{4[Q(g) - g \cdot P(g+1)] + (P(g+1))^2}}{2} \quad (9)$$

For the example in Table 1, we can thus determine h_r and g_r :

$$h_r = \frac{(4+1) \times 5 - 4 \times 3}{1 - 3 + 5} = \frac{13}{3} \approx 4.33$$

$$g_r = \frac{0 + \sqrt{4(30 - 5 \times 0) + 0^2}}{2} = \frac{\sqrt{120}}{2} \approx 5.48$$

4. The relation between the rational and real indicators

The real and rational variants of the h - and g -index look superficially similar in that they both interpolate between h and $h + 1$, and between g and $g + 1$. This raises the question which, if any, relations exist between h_{rat} and h_r , and between g_{rat} and g_r . We recall that equations (3), (4), (6) and (9) are the defining equations of these four indices.

4.1. The relation between h_{rat} and h_r

In some cases, $h_{rat} \geq h_r$ while in others $h_{rat} \leq h_r$. The two inequalities are possible as shown by the examples in Table 2. In case A we have: $h = 1$, $h_r = 1.5$ and $h_{rat} = 4/3$, hence $h_{rat} < h_r$. In case B we have: $h = 1$, $h_r = 1.5$ and $h_{rat} = 5/3$, hence $h_r < h_{rat}$.

Table 2. Two examples illustrating possible relations between h_{rat} and h_r

Case A		Case B	
r	$P(r)$	r	$P(r)$
1	3	1	2
2	0	2	1

We first note that if $h = 0$ then $h = h_r = h_{rat}$. We assume now that $h > 0$.

Using equation (3) we first rewrite h as a function of h_{rat} :

$$h = h_{rat} + \frac{n}{2h+1} - 1 \quad (10)$$

From (6) we obtain

$$h \cdot P(h) + P(h) - h \cdot P(h+1) = h_r \cdot (P(h) - P(h+1) + 1) \quad (11)$$

Now, we distinguish two cases: $P(h) = P(h+1)$ and $P(h) > P(h+1)$.

Case I: $P(h) = P(h+1)$

In this case $h = P(h)$ (by equation (1)), and $P(h) = h_r$ (by equation (6)), hence $h = h_r$. Hence we see that the precise relation between h_r and h_{rat} , using equation (3), is here given as:

$$h_{rat} = h_r - \frac{n}{2h+1} + 1 \quad (12)$$

This equation shows that $h_r < h_{rat}$. Indeed, equality only happens if $P(h+1) = 0$, but then $P(h) = 0$, or $h = 0$, which is already excluded. An example of this case is presented in Table 3.

Table 3. An example where $P(h) = P(h+1) = h = h_r < h_{rat} = 4/3$

r	$P(r)$
1	1
2	1

Case II: $P(h) > P(h+1)$

Now, we derive from (11) that $h \cdot (P(h) - P(h+1)) = h_r \cdot (P(h) - P(h+1) + 1) - P(h)$.

Hence: $h = h_r + \frac{h_r - P(h)}{P(h) - P(h+1)}$.

Substituting equation (10) leads to the precise relation:

$$h_{rat} = h_r + \frac{h_r - P(h)}{P(h) - P(h+1)} - \frac{n}{2h+1} + 1 \quad (13)$$

As $P(h) > P(h+1)$, the denominator of the first fraction is always positive. If now, moreover, $h_r \geq P(h)$ then clearly $h_{rat} > h_r$. This is not always the case, as shown by Table 2, case A, but in practical situations (author or journal citations) it is often the case that $h = P(h)$, and hence that $h_r = h = P(h)$. Only if $P(h) > h \geq P(h+1)$ it may happen (but not always, as illustrated in Table 2) that $h_{rat} < h_r$. Hence, in a publication-citation context it is often the case that $h \leq h_r \leq h_{rat} < h + 1$.

4.2. The relation between g_{rat} and g_r

The relation between g_{rat} and g_r can be found analogously. We assume that $g > 0$. Equation (9) can be rewritten as:

$$\left(g_r - \frac{P(g+1)}{2}\right)^2 = \frac{4[Q(g) - g \cdot P(g+1)] + (P(g+1))^2}{4}$$

or:

$$Q(g) + \frac{(P(g+1))^2}{4} - \left(g_r - \frac{P(g+1)}{2}\right)^2 = g \cdot P(g+1)$$

or:

$$Q(g) - g_r^2 + g_r \cdot P(g+1) = g \cdot P(g+1) \quad (14)$$

We distinguish between two cases: either $P(g+1) = 0$, implying that $Q(g) = Q(g+1)$, or $P(g+1) > 0$, implying that $Q(g) < Q(g+1)$. Both cases are graphically represented in Figure 2.

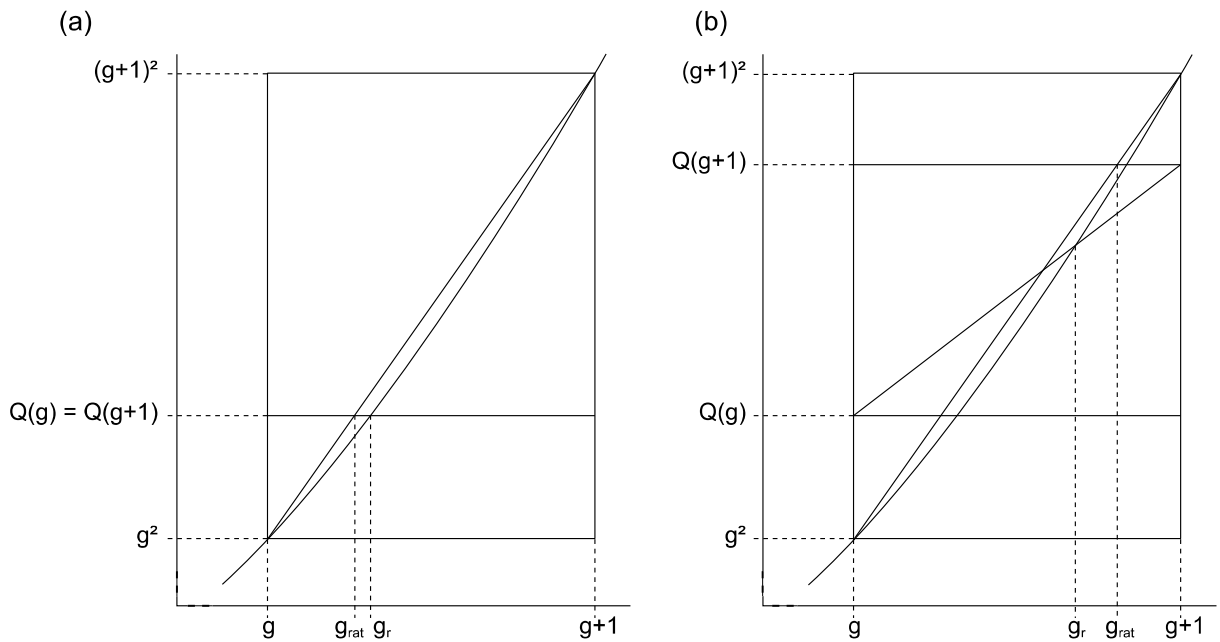


Figure 2. Graphical representation of g_{rat} and g_r , (a) in the case where $P(g+1) = 0$, and (b) in the case where $P(g+1) > 0$

Case I: $P(g+1) = 0$ or $Q(g) = Q(g+1)$

In this case $g_r = \sqrt{Q(g)}$ and $g_{rat} = g + \frac{Q(g) - g^2}{2g+1}$. From Figure 2a, it is obvious that this case always entails that $g_{rat} \leq g_r$. We provide a formal proof.

Theorem 1. If $P(g+1) = 0$, then $g_{rat} \leq g_r$.

Proof.

We first note that if $Q(g) = g^2$ and $P(g+1) = 0$, then $g = g_{rat} = g_r$. Suppose now that $Q(g) > g^2$. We have to prove that

$$g+1 - \frac{(g+1)^2 - Q(g)}{2g+1} < \sqrt{Q(g)}$$

This is equivalent to:

$$(g+1) \cdot (2g+1) < (2g+1) \cdot \sqrt{Q(g)} + (g+1)^2 - Q(g)$$

$$\text{or: } g^2 + g + Q(g) < (2g+1) \cdot \sqrt{Q(g)}$$

We consider this a quadratic inequality in the unknown $X = \sqrt{Q(g)}$ and have to show that:

$$X^2 - (2g+1) \cdot X + (g^2 + g) < 0 \text{ or } (X-g) \cdot (X-(g+1)) < 0$$

The inequality is proven if $g < \sqrt{Q(g)} < g+1$, which is indeed the case by equation (2). □

Case II: $P(g+1) > 0$ or $Q(g) < Q(g+1)$

From equation (14) we find that

$$g = g_r + \frac{Q(g) - g_r^2}{P(g+1)} \tag{15}$$

Combining (4) and (15) yields the relation between g_{rat} and g_r .

$$g_{rat} = g_r + \frac{Q(g) - g_r^2}{P(g+1)} + \frac{Q(g+1) - g^2}{2g+1} \tag{16}$$

In Figure 2b we see that $g_{rat} > g_r$. We will now prove that this is always the case in the discrete natural setting, i.e. where $P(x) \in \mathbf{N}$.

Theorem 2. If $P(g+1) > 0$, then $g_r < g_{rat}$.

Proof.

Given a certain g and $Q(g)$ ($< Q(g+1)$), g_r is a declining function of $P(g+1)$. It is therefore sufficient to prove the inequality for $P(g+1) = 1$.

We will henceforth denote $Q(g+1)$ as Q and $Q(g)$ as $Q-1$.

$$\text{Now } g_{rat} = g + 1 - \frac{(g+1)^2 - Q}{2g+1} = \frac{2g^2 + 2g + g + 1 - g^2 - 2g - 1 + Q}{2g+1} = \frac{g^2 + g + Q}{2g+1}$$

$$\text{while } g_r = \frac{1 + \sqrt{4(Q-g) - 3}}{2}.$$

Hence, we have to show that $\frac{g^2 + g + Q}{2g+1} > \frac{1 + \sqrt{4(Q-g) - 3}}{2}$, or

$$\frac{g^2 + g + Q}{2g+1} - \frac{1}{2} > \frac{\sqrt{4(Q-g) - 3}}{2},$$

or

$$2g^2 + 2Q - 1 > (2g+1) \cdot \sqrt{4(Q-g) - 3}.$$

Squaring both sides leads to:

$$4g^4 + 4Q^2 + 1 + 8g^2Q - 4g^2 - 4Q > (4g^2 + 4g + 1) \cdot (4Q - 4g - 3)$$

Now, $Q = g^2 + x$, where $1 \leq x \leq 2g$. Rewriting Q as $g^2 + x$ leads to:

$$16g^4 + 16g^2x + 4x^2 - 4x - 8g^2 + 1 > 16g^4 + 16g^2x - 24g^2 + 16gx - 16g + 4x - 3$$

or:

$$16g - 8x + 16g^2 - 16gx + 4x^2 + 4 > 0$$

This can be simplified to:

$$4g - 2x + (2g - x)^2 + 1 > 0$$

As $1 \leq x \leq 2g$, we have $2x \leq 4g$, hence $4g + (2g - x)^2 + 1 > 2x$. □

In summary, we do find some relations between the rational and real indicators but these are different depending on the precise situation. An overview can be found in Tables 4 and 5.

Table 4. Overview of the possible relations between h , h_r and h_{rat}

$h = 0$	$h > 0$	
	$P(h) = P(h+1)$	$P(h) > P(h+1)$
$h = h_r = h_{rat} = 0 = g$	$P(h) = h = h_r < h_{rat}$	All cases are possible.

Table 5. Overview of the possible relations between g , g_r and g_{rat}

$g = 0$	$g > 0$		
	$P(g+1) = 0$		$P(g+1) > 0$
	$Q(g) = g^2$	$Q(g) > g^2$	
$g = g_r = g_{rat} = 0 = h$	$g = g_r = g_{rat}$	$g_{rat} < g_r = \sqrt{Q(g)}$	$g_r < g_{rat}$

5. Maximum differences between h , h_r and h_{rat} , and between g , g_r and g_{rat}

In the previous section we have studied the relation, in the sense of being larger or smaller between h , h_r and h_{rat} on the one hand and between g , g_r and g_{rat} on the other. In this section we study the maximum difference between the h -type indices and between the g -type indices. Of course, by definition, this difference is always smaller than 1. In all cases the maximum difference can be as close to 1 as one wants (at least theoretically), simply by increasing h or g .

a) Maximum difference between h and h_{rat}

From the defining equation (3) it is clear that, for fixed h , this maximum difference is equal to $\frac{2h}{2h+1}$. This also means that for variable h this difference can be as close to 1 as one wants. An example of such a case is presented in Table 6, column A.

Table 6. Examples of the maximum difference between h , h_r and h_{rat}

	A	B	C
Rank r	$P(r)$	$P(r)$	$P(r)$
1	$h + 1$	L	$h + 1$
2	$h + 1$	L	$h + 1$
...
h	$h + 1$	L	h
$h + 1$	h	h	h

b) Maximum difference between h and h_r

This maximum difference can, even for fixed h , be as close to 1 as one wants (again in theory). An example is provided in Table 6, column B, where L denotes a large

number. In this example h_r is, by equation (6), equal to $\frac{L \cdot (h+1) - h^2}{1 - h + L}$. Hence the difference is equal to $\frac{L - h}{L - h + 1}$, which can, by increasing L , be made as close to one as one wants.

c) Maximum difference between h_{rat} and h_r

Taking h_r as small as possible, namely equal to h , and h_{rat} as large as possible, under the circumstances yields the largest possible difference equal to $\frac{2h-1}{2h+1}$. This is illustrated in Table 6, column C. If h is variable this difference can be made as close to 1 as one wants. Note that when putting $P(h) = h + 1$ (and not h) yields $h_r = h + \frac{1}{2}$, which leads to a much smaller difference.

d) Maximum difference between g and g_{rat}

From the defining equation (4) it is clear that, for fixed g , this maximum difference is equal to $\frac{2g}{2g+1}$. For variable g this difference can be as close to 1 as one wants. An example of such a case is presented in Table 7, column D.

Table 7. Examples of the maximum difference between g , g_r and g_{rat}

	D	E	F
Rank r	$Q(r)$	$Q(r)$	$Q(r)$
1			
2			
...
g	$g^2 + a$ ($0 \leq a \leq 2g$)	$g^2 + 2g$	g^2
$g + 1$	$g^2 + 2g$	$g^2 + 2g$	$g^2 + 2g$

e) Maximum difference between g and g_r

For fixed g , the difference between g and g_r is maximal if we take $Q(g)$ as large as possible, namely equal to $g^2 + 2g$. It automatically follows from formula (2) that in this

case $P(g+1) = 0$. Then, by equation (9), $g_r = \frac{\sqrt{4(g^2 + 2g)}}{2} = \sqrt{g^2 + 2g}$. The maximum difference is equal to $\sqrt{g^2 + 2g} - g = \frac{2}{\sqrt{1 + \frac{2}{g}} + 1}$. For variable g this expression can

be as close to 1 as one wants. An example of such a case is presented in Table 7, column E.

f) Maximum difference between g_{rat} and g_r

If $Q(g) = g^2$, then $g_r = \frac{P(g+1) + \sqrt{[2g - P(g+1)]^2}}{2} = g$. The maximum difference

between g_{rat} and g_r for fixed g is then equal to $\frac{2g}{2g+1}$, similar to case (d). Again, for

variable g this difference can be as close to 1 as one wants. An example of this case is presented in Table 7, column F. Note that column F is just a stricter variation of column D.

6. Conclusions

We reviewed the rational and real-valued variants of the h -index and g -index. These two indices are interesting additions to the standard constructions in that they interpolate between h and $h + 1$ or between g and $g + 1$. Their values can be obtained graphically as well as by a mathematical formula.

While there is no universally true relation between these variants, it is shown that in a citation context the real h -index is often smaller than the rational h -index. Furthermore, in a citation context it is rare for the publication ranked $g + 1$ to have no citations. We can therefore conclude that also for the g -index it is true that the real variant is typically smaller than the rational one. The maximum differences between h , h_r and h_{rat} on the one hand and between g , g_r and g_{rat} on the other are always close to one and grow larger with increasing h or g .

These two variants take the number of publications at the next position ($h+1$ or $g+1$) into account. There is, however, one exception: if $h = P(h)$ then $h_r = h$, whatever the value of $P(h+1)$ and similarly, if $g^2 = Q(g)$ then $g_r = g$, again independent of $P(g+1)$. In those cases the real-valued variants are always smaller than or equal than the rational variants. These cases, moreover, lead to the largest differences between the real-valued and the rational variants.

References

- Chai, J.C., Hua, P.H., Rousseau, R. and Wan, J.K. (2008). The adapted pure h -index. In: *Proceedings of WIS 2008, Berlin. Fourth International Conference on Webometrics, Informetrics and Scientometrics & Ninth COLLNET Meeting* (H. Kretschmer and F. Havemann, eds.).
<http://www.collnet.de/Berlin-2008/ChaiWIS2008aph.pdf>
- Egghe, L. (2006a). An improvement of the h -index: The g -index. *ISSI Newsletter*, 2(1): 8–9.
- Egghe, L. (2006b). Theory and practice of the g -index. *Scientometrics*, 69(1): 131–152.
- Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences*, 102(46): 16569–16572.
- Rousseau, R. (2006). Simple models and the corresponding h - and g -index. E-LIS: ID 6153, <http://eprints.rclis.org/archive/00006153/>.
- Ruane, F. and Tol, R.S.J. (2008). Rational (successive) h -indices: An application to economics in the Republic of Ireland. *Scientometrics*, 75(2): 395–405.
- Tol, R.S.J. (2008). A rational, successive g -index applied to economics departments in Ireland. *Journal of Informetrics*, 2(2): 149–155.