# A citation index bridging Hirsch's *h* and Egghe's *g*

Ruheyan Nuermaimaiti [a,b,*], Leonid Bogachev [a] and Jochen Voss [a]

[a]Department of Statistics, School of Mathematics, University of Leeds, Woodhouse Lane, Leeds LS2 9JT, United Kingdom
[b]Department of Mathematics, Faculty of Natural Sciences, Imperial College London, South Kensington Campus, London SW7 2AZ, United Kingdom
*To whom correspondence should be addressed: Email: R.Nuermaimaiti@imperial.ac.uk
**Edited By** Derek Abbott

## Abstract

We propose a new citation index $v$ ("nu") and show that it lies between the classical $h$-index and $g$-index. This idea is then generalized to a monotone parametric family $(v_a)$ $(a \geq 0)$, whereby $h = v_0$ and $v = v_1$, while the limiting value $v_\infty$ is expressed in terms of the maximum citation.

**Keywords:** citation indexes, $h$-index, $g$-index, scientometrics, citation data

> ### Significance Statement
>
> The widely used Hirsch's $h$-index values productivity but overlooks how highly each paper is cited, while Egghe's $g$-index emphasizes top-cited work but neglects lower-cited contributions. To address these imbalances, we propose the $v$-index, a synthetic metric that accounts for both highly and modestly cited publications and, therefore, offers a fairer and more balanced assessment of research impact.

## Introduction

### Background

Hirsch (1) made a breakthrough in scientometrics by proposing for the first time a simple citation index (commonly referred to as *h-index*), which had the advantage of aggregating the author's productivity on the basis of both the number of published papers and their quality measured by generated citations. Before that, only some extensive summary statistics were used, such as the mean number of citations per paper. Since then, the *h-index* has become a standard metric of authors' reputation and productivity, for instance routinely taken into consideration in academic appointments and promotions.

Specifically, the $h$-index is defined as the maximum number $h$ of an author's papers, each cited at least $h$ times (1). Therefore, this index only takes into account the fact of a relatively "high" citation of a paper, but the actual number of citations of such a paper is effectively ignored.

To remedy such censoring of larger citations, an alternative citation index (referred to as *g-index*) was proposed by Egghe (2), defined as the maximum number $g$ of an author's most cited papers, such that their total number of citations is at least $g^2$. From this definition, it is easy to see that $h \leq g$ (3).

These two (by now classical) indexes have attracted a lot of interest and generated ample research into their analytic properties and performance on real datasets, including their estimation in a variety of statistical models of count data (see e.g. (3–6)). Furthermore, many modifications and alternative variants of

the *h* and *g* indexes have been proposed, focusing on certain features of the citation profile (see e.g. (6–9) and further references therein).

### New index and layout

In the present work, we introduce a new citation index $v$ ("nu") aiming to bridge the mathematical definitions of Hirsch's $h$ and Egghe's $g$. This idea was first coined in Ref. (10). Namely, we start by observing that the *h-index* can be represented as a sum of certain indicator functions that censor papers to ensure a required minimum of citations. Building on this observation, our *v-index* essentially mimics the summative nature of the *h-index* but the new summation explicitly involves the numbers of citations of the top papers.

We are then able to show that our $v$ is "sandwiched" between $h$ and $g$; more precisely, we prove the two-sided inequalities

$$h \leq v \leq g^*,$$

where $g^*$ denotes a modified (unconstrained) $g$-index, obtained if we are allowed to add fictitious zeros to the citation vector (11). On the other hand, a "tempered" version $\bar{v}$ of the $v$-index, modified so as to be not larger than the number of published papers, satisfies the inequalities

$$h \leq \bar{v} \leq g.$$

We will finish off by introducing a more general family of citation indexes $(v_a)$, where $v_a$ is a non-decreasing (integer-valued)

function of a real parameter $\alpha \geq 0$. Here, $h = v_0$ and $v = v_1$, while the limiting value $v_\infty$ can be expressed in terms of the maximum citation.

## Disclaimer

There have been a lot of discussions about the utility and limitations of the indexes $h$ and $g$ (see e.g. (6, 7, 12–15) and references therein), including their questionable predictive power. In this work, we only interpret citation indexes as a suitable characteristic of productivity. However, some further thoughts about the societal dimension of citation indexes will be added in the Conclusion section.

## The $h$ and $g$ indexes

### Notation

Let us fix some notation. Suppose that an author has published $m \geq 1$ papers, with the ordered numbers of citations $x_1 \geq \ldots \geq x_m$, where $x_i \geq 0$ are integers (possibly zero). We call $\boldsymbol{x} = (x_1, \ldots, x_m)$ the *citation vector*. A zero vector $\boldsymbol{0} = (0, \ldots, 0)$ represents the degenerate case of no citations (note that the variable dimension of this vector is determined by the number of published papers). Denote

$$S_k = x_1 + \ldots + x_k, \quad k = 1, \ldots, m. \tag{1}$$

Clearly, $n = S_m$ is the total number of citations generated by the $m$ papers. Furthermore, we write

$$m_*(j) = \sum_{i=1}^{m} \mathbf{1}_{\{x_i \geq j\}} = \#\{x_i \geq j\} \tag{2}$$

for the number of papers with at least $j$ citations each; here, $\mathbf{1}_A = 1$ if condition $A$ is satisfied and $\mathbf{1}_A = 0$ otherwise.

Following (11), we say that a vector $\boldsymbol{x} = (x_1, \ldots, x_m)$ is *dominated* by a vector $\boldsymbol{y} = (y_1, \ldots, y_\ell)$ (written as $\boldsymbol{x} \preccurlyeq \boldsymbol{y}$) if $x_i \leq y_i$ for all $i \geq 1$; more precisely, if $m \leq \ell$ then $x_i \leq y_i$ for $i \leq m$, but if $m > \ell$ then $x_i \leq y_i$ for $i \leq \ell$ and $x_i = 0$ for $\ell < i \leq m$. Effectively, these two cases imply that we complement the absent components of either $\boldsymbol{x}$ or $\boldsymbol{y}$ with fictitious zeros to equalize their dimensions, and then the dominance holds component-wise.

REMARK 1 The component-wise dominance $\boldsymbol{x} \preccurlyeq \boldsymbol{y}$ should not be confused with *(weak) majorization* $\boldsymbol{x} \prec_w \boldsymbol{y}$, defined by the conditions $\sum_{i=1}^{k} x_i \leq \sum_{i=1}^{k} y_i$, for all $k$ (see Ref. (16), p. 11–12). Like before, the lengths of vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ are equalized by adding fictitious zeros as necessary. Clearly, if $\boldsymbol{x} \preccurlyeq \boldsymbol{y}$ then $\boldsymbol{x} \prec_w \boldsymbol{y}$, but not conversely.

### Generic properties

The following natural conditions are commonly assumed for any reasonable citation index $c(\boldsymbol{x})$ (11) (cf. (17)):

(C1) If $\boldsymbol{x} = \boldsymbol{0}$ then $c(\boldsymbol{x}) = 0$.
(C2) If $\boldsymbol{x} = (x_1, \ldots, x_m)$ and $\boldsymbol{y} = (x_1, \ldots, x_m, 0)$ then $c(\boldsymbol{x}) = c(\boldsymbol{y})$.
(C3) If $\boldsymbol{x} \preccurlyeq \boldsymbol{y}$ then $c(\boldsymbol{x}) \leq c(\boldsymbol{y})$.

REMARK 2 The majorization relation $\prec_w$ looks more flexible as a comparative tool. The corresponding version of property (C3) is stated similarly:

(C3) ′ If $\boldsymbol{x} \prec_w \boldsymbol{y}$ then $c(\boldsymbol{x}) \leq c(\boldsymbol{y})$.

However (perhaps, surprisingly), the *h-index* does not satisfy (C3′): e.g. for $\boldsymbol{x} = (2, 2)$ and $\boldsymbol{y} = (8, 1)$ we have $\boldsymbol{x} \prec_w \boldsymbol{y}$ but $h(\boldsymbol{x}) = 2 > h(\boldsymbol{y}) = 1$.

On the other hand, the $g^*$-index (but not $g$) does satisfy (C3′) (see definitions Eq. 6 and Eq. 8 below); e.g. $g^*(\boldsymbol{x}) = 2 < g^*(\boldsymbol{y}) = 3$.

## Mathematical expressions and relations for $h$ and $g$

Let us now recall the above verbal definitions of the $h$ and $g$ indexes and put them into an explicit mathematical formulation. Starting with the *h-index*, its definition can be expressed as follows,

$$h \equiv h(\boldsymbol{x}) = \max\left\{ j \geq 1 : \sum_{i=1}^{m} \mathbf{1}_{\{x_i \geq j\}} \geq j \right\}, \tag{3}$$

or, using notation Eq. 2,

$$h = \max\{ j \geq 1 : m_*(j) \geq j \}. \tag{4}$$

Note that the maximum in Eq. 3 is uniquely defined, since the sum on the left-hand side is a decreasing function of $j$, while the right-hand side of the testing inequality is strictly increasing.

In particular, noting that $m_*(j) \leq m$, it follows that the *h-index* is bounded by the number of papers:

$$h \leq m.$$

In the degenerate case with $x_i \equiv 0$ (i.e. $\boldsymbol{x} = \boldsymbol{0}$), the inequality in Eq. 3 is only satisfied for the value $j = 0$, which is excluded from the testing range; thus, the resulting set of suitable $j$'s is empty and, according to the common convention, its maximum is set to be zero: $\max \emptyset = 0$; hence $h(\boldsymbol{0}) = 0$, so that property (C1) is automatically satisfied. It is also easy to see that (C2) holds (because $h$ is insensitive to zero citations) and that (C3) is also true.

Next, the definition of the $g$-index can be written as follows,

$$g \equiv g(\boldsymbol{x}) = \max\left\{ 1 \leq k \leq m : \sum_{i=1}^{k} x_i \geq k^2 \right\}, \tag{5}$$

or, recalling notation Eq. 1,

$$g = \max\{ 1 \leq k \leq m : S_k \geq k^2 \}. \tag{6}$$

Note that if $\boldsymbol{x} = \boldsymbol{0}$ (i.e. all $x_i = 0$) then the set under the max-symbol is empty, in which case, by the same convention, we define the maximum as zero. That is to say, the $g$-index for the zero citation vector equals zero:

$$g(\boldsymbol{0}) = 0. \tag{7}$$

Also note that, because the testing range of $k$'s in Eq. 6 is bounded by $m$, we must have

$$g \leq m.$$

It can be shown that the $g$-index is not smaller than the *h-index* of the same author ((3), Proposition I.2. p. 133),

$$h \leq g.$$

Indeed, if the *h-index* has value $h$ then there are $h$ papers with at least $h$ citations each, and therefore with at least $h \times h = h^2$ citations in total. Hence, the trial value $k = h$ satisfies the inequality condition in Eq. 6, which implies that $g \geq k = h$, as claimed.

## Auxiliary lemmas for sums

According to definition Eq. 6, the index $g$ is the largest value of $k \leq m$ for which $S_k \geq k^2$. But it may be unclear whether the inequality $S_k \geq k^2$ can fail for some $k < g$. Let us show that $S_k \geq k^2$ *for all $k \leq g$*.

LEMMA 1 If $S_k < k^2$ for some $k \geq 1$, then $S_\ell < \ell^2$ for all $\ell \geq k$.

*Proof.* Using that $x_k = \min\{x_1, \ldots, x_k\}$, we have

$$k^2 > S_k = x_1 + \cdots + x_k \geq kx_k,$$

which implies that $k > x_k \geq x_{k+1}$. Hence,

$$S_{k+1} = S_k + x_{k+1} < k^2 + k < (k+1)^2,$$

that is, $S_{k+1} < (k+1)^2$. The general claim then follows by induction.

LEMMA 2    If $S_k < k^2$ then $g < k$. In particular, $g < m$ or $g = m$ according as $S_m < m^2$ or $S_m \geq m^2$, respectively.

*Proof.*    Readily follows by Lemma 1 and definition Eq. 6.

## The unconstrained index $g^*$

Turning to the verification of the required properties (C1)–(C3) for the $g$-index, we see that (C1) automatically holds due to Eq. 7. It is also easy to see that (C3) holds as well. However, the result of Lemma 2 suggests, a bit surprisingly, that property (C2) may fail. For instance, for $\mathbf{x} = (4)$ we have $g(\mathbf{x}) = 1$, but for $\mathbf{y} = (4, 0)$ definition Eq. 6 yields $g(\mathbf{y}) = 2$.

To salvage (C2), and also to amplify the role of top-cited papers, it was suggested (3, 11) to lift the constraint $k \leq m$ in the definition of the $g$-index (see Eq. 6) by complementing the citation vector $\mathbf{x} = (x_1, \ldots, x_m)$ with additional zeros, as if such fictitious papers have been published but generated no citations: $\mathbf{x}' = (x_i') = (x_1, \ldots, x_m, 0, \ldots)$. We denote this version of $g$ by $g^*$:

$$g^* \equiv g^*(\mathbf{x}) = \max\left\{ k \geq 1 : \sum_{i=1}^{k} x_i' \geq k^2 \right\} \qquad (8)$$

or, equivalently,

$$g^* = \max\{ k \geq 1 : S_k \geq k^2 \}, \qquad (9)$$

where we define $S_k = S_m$ for all $k \geq m$.

Comparing definitions Eq. 6 and Eq. 8, we see that

$$g \leq g^*,$$

and moreover, if $g < m$ then $g = g^*$. However, the case where $g = m$ may be drastically different.

LEMMA 3    Suppose that $S_m \geq (m+1)^2$. Then $g = m$ but $g^* = \lfloor \sqrt{S_m} \rfloor \geq m + 1$.

*Proof.*    Note that $g = m$ by Lemma 2. By definition Eq. 9, we have

$$S_{g^*} = S_m \geq (g^*)^2, \quad S_{g^*+1} = S_m < (g^*+1)^2.$$

In turn, this implies

$$\sqrt{S_m} - 1 < g^* \leq \sqrt{S_m},$$

that is, $g^* = \lfloor \sqrt{S_m} \rfloor \geq m + 1$, as claimed.

For example, for $\mathbf{x} = (5, 4)$ we have $m = 2$, $S_m = 9$, $g = 2$, and $g^* = 3$. A striking real-life example illustrating this situation is the case of John Nash (see Ref. (11)), with the (rounded) citation vector $\mathbf{x} = (2,000, 2,000, 1,500, 1,000, 400, 250, 100, 100)$, for which we get $g = 8$ but $g^* = 85$.

## An alternative citation index $v$
### Idea and definitions

Trying to reconcile the definition of the *h-index* given by formula (3), with the definition of the *g-index* in Eq. 5 by taking into account the actual citations of the top papers, we propose a new citation index called the *v-index*, defined as the maximum integer $v$ such that the total sum of citation counts of papers with at least $v$ citations each is not less than $v^2$. Mathematically, this is expressed as (cf. Eq. 3)

$$v \equiv v(\mathbf{x}) = \max\left\{ j \geq 1 : \sum_{i=1}^{m} x_i \mathbf{1}_{\{x_i \geq j\}} \geq j^2 \right\}, \qquad (10)$$

or, equivalently,

$$v = \max\left\{ j \geq 1 : S_{m_*(j)} \geq j^2 \right\}. \qquad (11)$$

Similarly to Eq. 3, the maximum is uniquely defined, noting that the sum in Eq. 10 is a decreasing function of $j$, while the right-hand side is strictly increasing. It is also worth pointing out that, unlike $g$ vs. $g^*$, the $v$-index is insensitive to fictitious zeros.

Simple examples show that the value of $v$ may be larger than the total number of papers, in contrast with the $h$ and $g$ indexes. For instance, for $\mathbf{x} = (9, 7, 1)$ we get $v(\mathbf{x}) = 4 > 3$.

Clearly, this occurs because our definition of $v$ gives prominence to few highly cited papers. If unwanted, this can be suppressed by modifying the definition via an explicit constraint $v \leq m$:

$$\bar{v} \equiv \bar{v}(\mathbf{x}) = \max\left\{ 1 \leq j \leq m : \sum_{i=1}^{m} x_i \mathbf{1}_{\{x_i \geq j\}} \geq j^2 \right\}, \qquad (12)$$

or, equivalently,

$$\bar{v} = \max\left\{ 1 \leq j \leq m : S_{m_*(j)} \geq j^2 \right\}.$$

We call $\bar{v}$ a *tempered v-index*.

## Checking the basic properties

LEMMA 4    The indexes $v$ and $\bar{v}$ satisfy the basic properties (C1)–(C3).

*Proof.*    Properties (C1) and (C2) are straightforward, since $v$ and $\bar{v}$ are insensitive to zero values $x_i = 0$. Monotonicity (C3) is also obvious because the sums in Eqs. 10 and 12 are monotone increasing in each component $x_i$.

The $v$-index combines the features of both the *h-index* and the *g-index*. It takes into account citations that are equal to or greater than a minimum threshold value of $v$ as in the *h-index*, while also including higher citations as in the *g*-index. This ensures that the $v$-index captures the impact of highly cited papers and provides a more balanced picture of their overall scholarly impact. In particular, it may be expected that the $v$-index interpolates between $h$ and $g$. The next result supports this conjecture.

## Main result—ordering relations between the indexes

THEOREM 1    The citation indexes $h, v, \bar{v}, g$, and $g^*$ are in the following ordering relations:

$$h \leq v \leq g^*, \quad h \leq \bar{v} \leq g. \qquad (13)$$

*Proof.*    We only prove the inequalities for $v$; the proof for $\bar{v}$ is similar. First, by definition of $h$ in Eqs. 3 and 4, we can write

$$h \leq m_*(h) = \sum_{i=1}^{m} \mathbf{1}_{\{x_i \geq h\}} \leq \frac{1}{h}\sum_{i=1}^{m} x_i \mathbf{1}_{\{x_i \geq h\}} = \frac{1}{h}S_{m_*(h)}.$$

Hence,

$$S_{m_*(h)} = \sum_{i=1}^{m} x_i \mathbf{1}_{\{x_i \geq h\}} \geq h^2,$$

which implies, according to Eq. 11, that $v \geq h$, as claimed.

Next, the maximizing sum in Eq. 10 is expressed as

$$v^2 \leq \sum_{i=1}^{m} x_i \mathbf{1}_{\{x_i \geq v\}} = S_{m_*(v)}, \tag{14}$$

thus involving $m_*(v)$ terms $x_i$ satisfying the inequality

$$x_i \geq v, \quad i = 1, \ldots, m_*(v). \tag{15}$$

If $m_*(v) \leq v$ then from Eq. 14 we obtain (adding fictitious zeros if $v > m$)

$$v^2 \leq S_{m_*(v)} \leq S_v,$$

and it follows from definition Eq. 9 that $g^* \geq v$. Alternatively, if $m_*(v) \geq v$ then, using Eq. 15, we can write

$$S_{m_*(v)} \geq S_v \geq v^2,$$

and, as before, it follows that $g^* \geq v$.

## R code and some simple examples

A simple R code to calculate various indexes is given below:

```
# indexes h, nu, nu.bar, g, g.star
  ind <- function(x) # x = input citation vector
  { x <- sort(x, decreasing = TRUE) # ordering
   m <- length(x) # number of papers
   # h
   h <- 0
   while (h < length(x) & x[h + 1] >= h + 1)
   { h <- h + 1
   }
   # nu
   nu <- 0
   while (sum(x[which(x >= (nu + 1))])
       >= (nu + 1)^2)
   { nu <- nu + 1
   }
   # nu.bar
   nu.bar <- min(nu,m)
   # g
   g <- max(which(cumsum(x) >= (1:m)^2))
   # g.star
   if (sum(x) >= m^2)
   { g.star <- floor(sqrt(sum(x)))
   }
   else
   { g.star <- max ( which (cumsum(x) >= (1:m)^2))
   }
  }
  # Printing the output:
  cat ("x =","(",x,");", "\n")
  cat ("h =", h, "nu.bar =", nu.bar, "nu =", nu,
     "g =", g, "g.star =", g.star)
```

Example (John Nash case):

```
x <- c(2000,2000,1500,1000,400,250,100,100)
  ind(x)
  # x = ( 2000 2000 1500 1000 400 250 100 100 );
  # h = 8 nu.bar = 8 nu = 85 g = 8 g.star = 85
```

The following Table 1 presents the various citation indexes for a few simple examples.

**Table 1.** Illustrative examples of different indexes.

| $\boldsymbol{x} = (x_1, \ldots, x_m)$ | $h$ | $\bar{v}$ | $v$ | $g$ | $g^*$ |
|---|---|---|---|---|---|
| (3, 2, 2, 2) | 2 | 2 | 2 | 2 | 2 |
| (12, 3, 1) | 2 | 3 | 3 | 3 | 4 |
| (12, 3, 1, 0) | 2 | 3 | 3 | 4 | 4 |
| (6, 3, 1, 0) | 2 | 3 | 3 | 3 | 3 |
| (5, 3, 2, 1) | 2 | 2 | 2 | 3 | 3 |
| (8, 1, 1) | 1 | 2 | 2 | 3 | 3 |
| (8, 4, 3, 2, 1) | 3 | 3 | 3 | 4 | 4 |
| (18, 18, 1, 1) | 2 | 4 | 6 | 4 | 6 |
| (20, 20, 18, 6, 1, 0) | 4 | 6 | 7 | 6 | 8 |

## Cases of equality

One observation from Table 1 is that, occasionally, some of the indexes may coincide, which warrants a question of exploring the cases of equalities in Eq. 13. The possible equality $h = g^*$ was addressed by Egghe et al. (18).

THEOREM 2  The equalities in the index inequalities Eq. 13 of Theorem 1 hold if and only if the following conditions are satisfied, respectively:

(a) $h = v$: $S_{m_*(h+1)} < (h+1)^2$;

(b) $v = g^*$: $S_{v+1} < (v+1)^2$;

(c) $h = \bar{v}$: $h = m$, or $h < m$ and $S_{m_*(h+1)} < (h+1)^2$;

(d) $v = g$: $v = m$, or $v < m$ and $S_{v+1} < (v+1)^2$.

*Proof.*  Since it is always true that $v \geq h$ (see Eq. 13), the equality $v = h$ simply means that $v < h + 1$. But, according to definition Eq. 11, the latter inequality is equivalent to $S_{m_*(h+1)} < (h+1)^2$, which is the claim of part (a). Essentially the same argument proves part (c), except that, due to the bound $\bar{v} \leq m$, a special case arises if $h = m$, which automatically implies $\bar{v} = m$.

Similarly, due to Eq. 13 we have $g^* \geq v$, while the inequality $g^* < v + 1$ is equivalent to $S_{v+1} < (v+1)^2$, according to definition Eq. 9, and the claim of part (b) follows. The same argument applies to part (d), with an additional consideration of the special case $v = m$.

Part (a) is exemplified by $\boldsymbol{x} = (3, 2, 1)$: here, $h = v = 2$, while $m_*(2) = 2$, $m_*(3) = 1$ and $S_2 = 5 > 2^2$ but $S_1 = 3 < 3^2$. The same example gives $g^* = 2$, confirmed by the inequality $S_3 = 6 < 3^2$, in line with part (b). Furthermore, since $g = g^*$ and $\bar{v} = v$, this example also illustrates parts (c) and (d). As for the boundary case of parts (c) and (d), it occurs, for example, for $\boldsymbol{x} = (4, 3, 3)$, where $h = \bar{v} = v = g = 3$. Another example of (b) is the John Nash case mentioned above.

## Indexes $h$, $v$, $g^*$
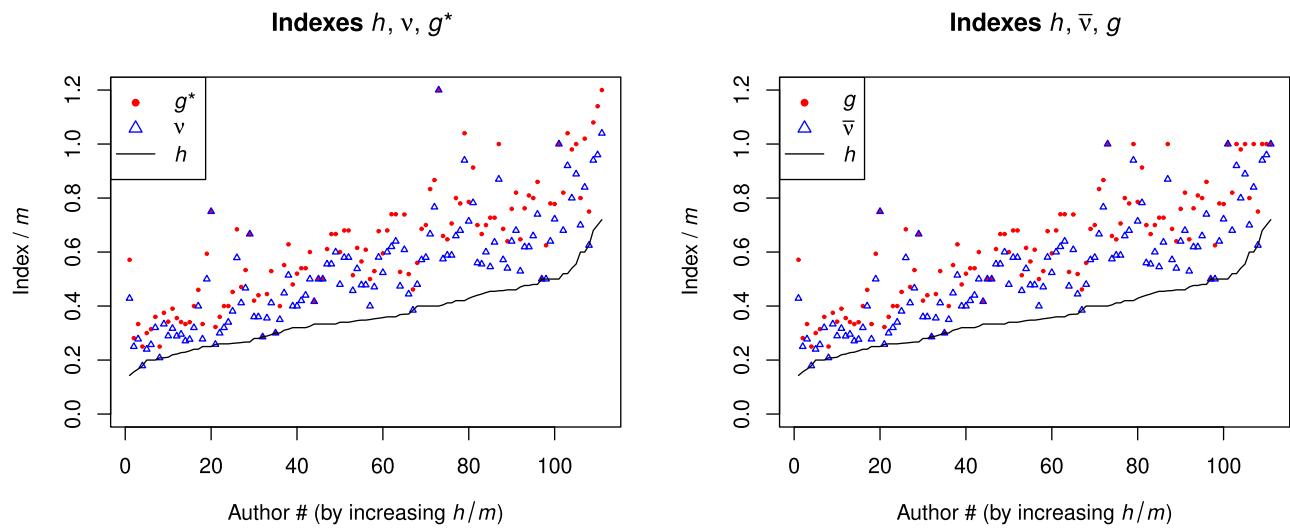


## Indexes $h$, $\bar{v}$, $g$

**Fig. 1.** Index plots for the EJP dataset, showing triplets of indexes ($h \leq v \leq g^*$ or $h \leq \bar{v} \leq g$) normalized by the number of published papers per author. The authors are ranked in increasing order with respect to the parameter $h/m$.

**Table 2.** Pairwise correlations across the citation indexes and the number of published papers ($m$).

| Index | $h$ | $v$ | $\bar{v}$ | $g$ | $g^*$ | $m$ |
|---|---|---|---|---|---|---|
| $h$ | 1.0000 | 0.9649 | 0.9646 | 0.9656 | 0.9725 | 0.8044 |
| $v$ | 0.9649 | 1.0000 | 0.9998 | 0.9932 | 0.9978 | 0.7743 |
| $\bar{v}$ | 0.9646 | 0.9998 | 1.0000 | 0.9942 | 0.9978 | 0.7768 |
| $g$ | 0.9656 | 0.9932 | 0.9942 | 1.0000 | 0.9967 | 0.8046 |
| $g^*$ | 0.9725 | 0.9978 | 0.9978 | 0.9967 | 1.0000 | 0.7893 |
| $m$ | 0.8044 | 0.7743 | 0.7768 | 0.8046 | 0.7893 | 1.0000 |

## Data example

Here, we illustrate the calculation of the various citation indexes for real data collected by the first-named author (available online at https://github.com/Ruheyan/WoS-citation-data).[a] The dataset comprises citation counts, with a cut-off date of 2022 September 19th, of 3,615 papers (with 73,730 citations in total) of 111 authors who published a paper in the first 10 issues of *Electronic Journal of Probability* (EJP), vol. 24 (2019) (https://projecteuclid.org/journals/electronic-journal-of-probability/volume-24/issue-none). The data were derived from the Web of Science ([20]).

Figure 1 shows the plots representing the indexes $h$, $\bar{v}$, $v$, $g$, and $g^*$ (in triplets, for ease of comparison) for all 111 authors, normalized by the number of papers per author. The calculated values confirm the inequalities of Theorem 1, but one can observe that the new index ($v$ or $\bar{v}$) tends to be closer to the upper bound $g^*$ or $g$, respectively. Furthermore, Table 2 presents correlations between different indexes—not surprisingly, they are all strongly positively correlated (especially in the "sister" pairs ($v$, $\bar{v}$) and ($g$, $g^*$)), but correlation with the number of papers ($m$) is weaker.

## Parametric family ($v_\alpha$)
### Definition and monotonicity

It is quite natural to generalize the definition of the index $v$ in Eq. 10 by considering different powers. Namely, for $\alpha \geq 0$ we define the $v_\alpha$-index as

$$v_\alpha \equiv v_\alpha(\mathbf{x}) = \max\left\{ j \geq 1 : \sum_{i=1}^{m} x_i^\alpha \mathbf{1}_{\{x_i \geq j\}} \geq j^{\alpha+1} \right\}. \quad (16)$$

Clearly, for $\alpha = 0$ and $\alpha = 1$ this definition is reduced to Eqs. 3 and 10, respectively:

$$v_0 = h, \quad v_1 = v.$$

Like in Eqs. 3 and 10, the existence and uniqueness of the maximum in Eq. 16 is self-evident, noting that the sum is a decreasing function of $j$ while the right-hand side is strictly increasing. It is straightforward to verify that $v_\alpha$ satisfies (C1)–(C3). We also observe the monotonicity of the family ($v_\alpha$).

THEOREM 3 The function $v_\alpha$ is increasing in $\alpha \geq 0$.

*Proof.* Rewrite Eq. 16 as

$$v_\alpha = \max\left\{ j \geq 1 : \sum_{i=1}^{m} \left(\frac{x_i}{j}\right)^\alpha \mathbf{1}_{\{x_i \geq j\}} \geq j \right\}, \quad (17)$$

and note that the sum in Eq. 17 is monotone increasing in $\alpha$, since $x_i/j \geq 1$.

As an illustration of sensitivity and fluidity of $v_\alpha$, in the John Nash case it is easy to check that, for example, for $\alpha = 0.5$ we have $v_{0.5} = 35$, compared to $v_1 = g^* = 85$. R code to calculate $v_\alpha$ is given below:

```
# nu.alpha
  x <- sort(x, decreasing=TRUE)
  nu.alpha <- function(alpha)
  { sapply(alpha, function(a)
    { nu <- 0
      while (sum((x[x >= (nu + 1)] / (nu + 1))^a)
          >= (nu + 1))
      { nu <- nu + 1
      }
      return(nu)
    })
  }
  # Plotting the output:
  curve(nu.alpha, col = "red", lwd = 2,
      xlim = c(0, max(x)+20), ylim = c(1, max(x)),
      xlab = expression(paste(alpha)),
      ylab = expression(paste(nu[alpha])),
      main = bquote(paste(bold("x "), "= (",
      .(toString(x)), ")")))
```

The next Fig. 2 illustrates the behavior of the function $v_\alpha$ for some examples from Table 1. The reader may also find it interesting to run this code on the citation data of John Nash.

## The limit as $\alpha \to \infty$

It is interesting to understand the meaning of the limiting value $v_\infty = \lim_{\alpha\to\infty} v_\alpha$.

THEOREM 4  For a citation vector $\mathbf{x} = (x_1, \ldots, x_m)$, denote by $\ell_1 = \sum_{i=1}^{m} \mathbf{1}_{\{x_i = x_1\}} \equiv m_*(x_1)$ the multiplicity of the top citation $x_1 = \max\{x_i,\ 1 \le i \le m\}$. Then

$$v_\infty(\mathbf{x}) = \begin{cases} x_1 - 1 & \text{if } \ell_1 < x_1, \\ x_1 & \text{if } \ell_1 \ge x_1. \end{cases} \qquad (18)$$

*Proof.*   Follows using Eq. 17 by noting that $(x_1/j)^\infty$ equals $\infty$, 1 or 0 according as $j < x_1$, $j = x_1$ or $j > x_1$, respectively.
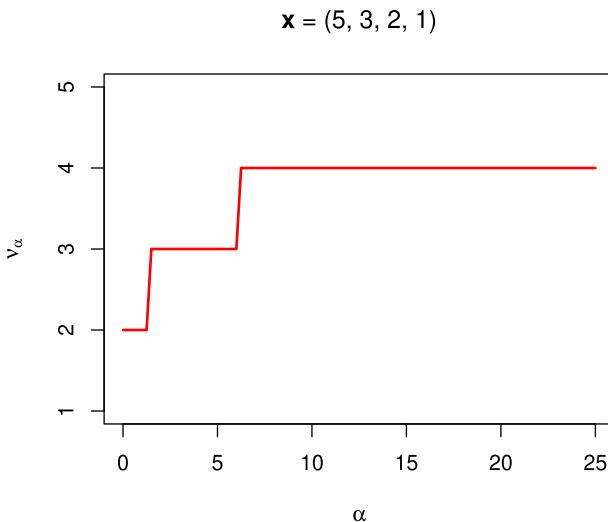
## Conclusion

We have introduced some new citation indexes starting with $v = v_1$, and investigated their relations with the classical indexes $h$ and $g$. As already mentioned, the *h-index* is straightforward and informative, but it is limited by only acknowledging the fact of a high citation but not the actual number of citations. In contrast, the *g-index* is based exclusively on the citations of a few top papers, but ignoring the "footing" of lower-cited papers.

Our synthetic proposal of the *v-index* is designed so as to take into account both higher and lower cited papers, which may assess the individual's productivity in a more fair and balanced way. Indeed, we have seen that the *v-index* is in a sense bridging Hirsch's *h* and Egghe's *g*. Furthermore, the spectrum of the indexes $(v_\alpha)$ provides a flexible toolkit that allows one either to enhance or to inhibit the input from top-cited papers, as required.

Of course, it goes without saying that none of these, or any other indexes known in the literature, is perfect and should replace the rest. In fact, a reasonable practical recommendation may be to choose a few indexes to judge someone's academic achievement, depending on the assessment requirements and also on the specific features of the scientific domain. In this regard, it may be useful to choose the parameter $\alpha$ in the index $v_\alpha$ according to certain individual features of the citation vector $\mathbf{x}$, in the spirit of limit theorems for norms of random vectors [21, 22]. We will address this issue in our future work.

In conclusion, we reiterate that prudence, maturity, and care should be exercised when using citation indexes in social practice, especially making sure to avoid misuse and/or abuse of their utility as predictors of future performance and productivity. Although citation indexes succinctly grasp some objective aggregated information from citation records, they are deceptively easy to compute, replacing individual research track records with a simple number, while these results should be verified and complemented by human evaluation by experts.

The scientometrics community has quickly realized, and extensively documented, the growing threat of misusing the *h-index* and other indicators for far reaching and often unjustified implications in the social interpretation (see e.g. [7, 12–14, 23] and further references therein). These concerns and wide discussions have led to the creation and promotion of good practice protocols,
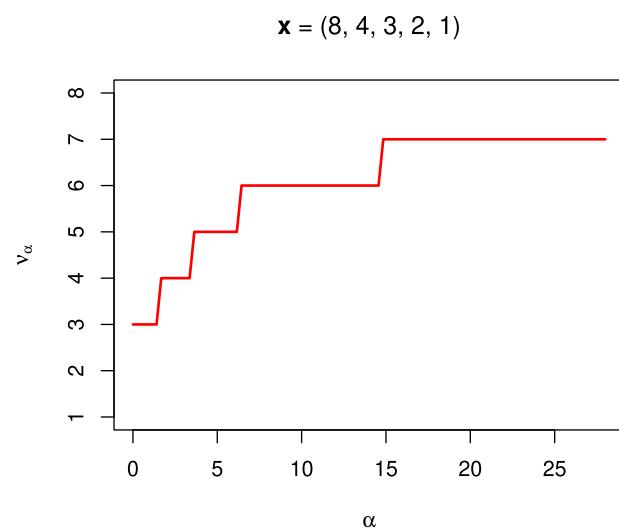
**Fig. 2.** Illustrative graphs of the index $v_\alpha$ as a function of parameter $\alpha \in [0, \infty)$. Note the values $v_0 = h$, $v_1 = v$, and $v_\infty = x_1 - 1$ (Eq. 18).

such as the San Francisco Declaration on Research Assessment (DORA) [23] or the Leiden Manifesto [24].

The risks are further amplified by the fast growing use of AI including Large Language Models such as ChatGPT, whereby the responsibility for conclusions and extrapolations may be delegated inadvertently to the computer [13]. Although deployment of AI for assistance in technical analyses and summarization is an inevitable and welcome trend, the best vaccine against misuse and abuse is to combine formal calculations and summaries with a robust comparison against the specific domain "golden standards," based on an objective expert evaluation and enhanced by a reproducible and unbiased statistical analysis.

## Note

[a] Similar citation data were used in a conference paper [19] and PhD thesis [10].

## Author Contributions

L.B. and R.N. designed the study and wrote the first draft of the article. L.B. and J.V. contributed to the mathematical derivation and interpretation of the results. All authors agreed with the final version of the article.

## Previous Presentation

These results were presented as a poster at the 29th Annual International Conference on Science and Technology Indicators (STI-ENID 2025), Bristol, UK, September 3–5, 2025 (https://virtual.oxfordabstracts.com/event/74626/submission/143).

## Preprints

A preprint with an earlier version of this article is published at https://arxiv.org/abs/2504.20600.

## Data Availability

EJP dataset used in this article is available online at https://github.com/Ruheyan/WoS-citation-data.

## References

1 Hirsch JE. 2005. An index to quantify an individual's scientific research output. *Proc Natl Acad Sci U S A*. 102(46):16569–16572.
2 Egghe L. 2006. An improvement of the h-index: the g-index. *ISSI Newsl*. 2(1):8–9.
3 Egghe L. 2006. Theory and practise of the *g*-index. *Scientometrics*. 69(1):131–152.
4 Egghe L, Rousseau R. 2006. An informetric model for the Hirsch-index. *Scientometrics*. 69(1):121–129.
5 Egghe L, Rousseau R. 2008. An *h*-index weighted by citation impact. *Inf Process Manag*. 44(2):770–780.
6 SCI²S. 2025. Soft Computing and Intelligent Information Systems. (Online) *h-Index and Variants*. University of Granada. https://sci2s.ugr.es/hindex.
7 Alonso S, Cabrerizo FJ, Herrera-Viedma E, Herrera F. 2009. *h*-index: a review focused in its variants, computation and standardization for different scientific fields. *J Informetr*. 3(4): 273–289.
8 Guns R, Rousseau R. 2009. Real and rational variants of the *h*-index and the *g*-index. *J Informetr*. 3(1):64–71.
9 Hirsch JE. 2019. $h_\alpha$: an index to quantify an individual's scientific leadership. *Scientometrics*. 118(2):673–686.
10 Nuermaimaiti R. 2023. Statistical models for frequency distributions of count data with applications to scientometrics [PhD thesis]. University of Leeds. https://etheses.whiterose.ac.uk/id/oai_id/oai:etheses.whiterose.ac.uk.33446.
11 Woeginger GJ. 2008. An axiomatic analysis of Egghe's *g*-index. *J Informetr*. 2(4):364–368.
12 Costas R, Bordons M. 2007. The h-index: advantages, limitations and its relation with other bibliometric indicators at the micro level. *J Informetr*. 1(3):193–203.
13 Thelwall M. 2025. Quantitative methods in research evaluation: citation indicators, altmetrics, and artificial intelligence [preprint], arXiv:2407.00135 [cs.DL] [https://doi.org/10.48550/arXiv.2407.00135]
14 Waltman L. 2016. A review of the literature on citation impact indicators. *J Informetr*. 10(2):365–391.
15 Hirsch JE. 2007. Does the h index have predictive power? *Proc Natl Acad Sci U S A*. 104(49):19193–19198.
16 Marshall AW, Olkin I, Arnold BC. 2011. *Inequalities: theory of majorization and its applications*. 2nd ed. Springer Series in Statistics. New York, Springer.
17 Woeginger GJ. 2008. An axiomatic characterization of the Hirsch-index. *Math Soc Sci*. 56(2):224–232.
18 Egghe L, Fassin Y, Rousseau R. 2019. Equalities between h-type indices and definitions of rational h-type indicators. *J Data Inform Sci*. 4(1):22–31.
19 Nuermaimaiti R, Bogachev LV, Voss J. A generalized power law model of citations. In: Glänzel W, Heeffer S, Chi P-S, Rousseau R, editors. *Proceedings, 18th International Conference on Scientometrics & Informetrics (ISSI 2021)*. International Society for Scientometrics and Informetrics (I.S.S.I.); 2021. p. 843–848. https://www.issi-society.org/proceedings/issi_2021/ProceedingsISSI 2021.pdf.
20 WoS. 2025. Web of Science. (Online) https://clarivate.com/academia-government/scientific-and-academic-research/research-discovery-and-referencing/web-of-science.
21 Bogachev L. 2006. Limit laws for norms of IID samples with Weibull tails. *J Theor Probab*. 19(4):849–873.
22 Schlather M. 2001. Limit distributions of norms of vectors of positive i.i.d. random variables. *Ann Probab*. 29(2):862–881.
23 DORA. 2020. San Francisco declaration of research assessment. American Society for Cell Biology (ASCB). (Online) https://sfdora.org/read.
24 Hicks D, Wouters P, Waltman L, de Rijcke S, Rafols I. 2015. Bibliometrics: the Leiden Manifesto for research metrics. *Nature*. 520(7548):429–431.