

Detecting *h*-index manipulation through self-citation analysis

Christoph Bartneck · Servaas Kokkelmans

Received: 15 July 2010 / Published online: 11 November 2010
© The Author(s) 2010. This article is published with open access at Springerlink.com

Abstract The *h*-index has received an enormous attention for being an indicator that measures the quality of researchers and organizations. We investigate to what degree authors can inflate their *h*-index through strategic self-citations with the help of a simulation. We extended Burrell's publication model with a procedure for placing self-citations, following three different strategies: random self-citation, recent self-citations and *h*-manipulating self-citations. The results show that authors can considerably inflate their *h*-index through self-citations. We propose the *q*-index as an indicator for how strategically an author has placed self-citations, and which serves as a tool to detect possible manipulation of the *h*-index. The results also show that the best strategy for an high *h*-index is publishing papers that are highly cited by others. The productivity has also a positive effect on the *h*-index.

Keywords *h*-Index · Self-citation · Manipulation · *q*-Index · Simulation

Introduction

In the competitive academic world, it is necessary to assess the quality of researchers and their organizations. The allocation of resources and individual careers depend on it. In the UK, for example, the use of bibliometric indicators for the national Research Assessment Exercise (RAE) has long been discussed [17]. Efforts have been made to make such assessments as objective as possible. While the productivity could relatively easy be measured by counting publications, assessing the impact of research often relied on the count of citations received. In 2005, Hirsch [9] proposed the *h*-index, which tries to bring productivity and impact into a balance.

C. Bartneck (✉)
Department of Industrial Design, Eindhoven University of Technology, Eindhoven, The Netherlands
e-mail: c.bartneck@tue.nl

S. Kokkelmans
Department of Physics, Eindhoven University of Technology, Eindhoven, The Netherlands
e-mail: s.kokkelmans@tue.nl

It is hard to underestimate the effect that his proposal had on the field of scientometrics. Google Scholar lists 1,130 citations for his original paper as of June 16th, 2010. Some authors even divide the research field into a pre and post Hirsch period [14]. A wealth of extensions, modifications have since been proposed [12, 13] and also new indicators have been developed, such as the g -index [4]. These modification and new indicators are intended to improve the original h -index. The arrival of the Publish and Perish software made the calculation of these diverse indicators accessible to a more general public.

An elaborated review on the benefits and problems of the h -index is available [5]. We will focus on the problem of self-citations, which has polarized the research community. On the one hand, self-citations can be considered a natural part of scientific communication, while others condemn it as a means to artificially inflate bibliometric indicators. Besides this fundamental divide about the role and importance of self-citations, there are also practical issues. Reliably filtering self-citations is currently only practical in highly consistent data sets, such as from the Web of Science. But even Thomson-Reuter had to introduce the Researcher ID to identify unique researchers, in particular if researchers have the same name. The Web of Science might be a useful data set for traditional disciplines, such as Physics, but its coverage is insufficient for research fields in which conference proceedings play an important role [11]. Google Scholar (GS) offers the widest coverage of academic communication, but filtering self-citations from its results is currently not reliably possible. And maybe we even should not filter them, since they form an organic part of the citation process [8] and self-citations make up for up to 36% of all citations [1]. This might make it difficult to sharpen the h -index by excluding self-citation as it was already proposed [15]. It has also been demonstrated that results from GS can potentially be manipulated through mock publication [10].

Still it would be useful to be able to distinguish between authors that cite their previous work to clarify the relationship with the paper at hand and authors that strategically cite their papers even if they are not directly relevant to the current paper.

One method of strategically manipulating the h -index is the following: first cite the paper(s) that have currently as many citations as the h -index and then proceed downwards from there. Lets look an example to illustrate this strategy. Figure 1 shows the citation profile for an example author that published 60 papers. His papers are sorted by the citations they have received. He currently has an h -index of 20, which is visualized by the diagonal line. This means that he has at least 20 papers that have each been cited at least 20 times. We will refer to the paper that has the least citations and still contributes to the h -index as the h -paper. In this case it is paper number 20. If he would cite the h -paper and paper number 21, that each have currently 20 citations, then his h -index would increase to 21. He would have 21 papers that each have at least 21 citations. With only investing two self-citations, this author could inflate his h -index by one. A more subtle strategy would be to only cite papers that currently have fewer citations than the author's h -paper since citing already highly cited papers is unlikely to increase the h -index quickly.

Given that up to 36% of all citations are self-citations, the potential inflation of bibliometric indicators could be enormous. We therefore focus on the following research questions:

- (1) How much can authors inflate their h -index through strategic self-citations?
- (2) How can we detect strategic self-citation?
- (3) What influence has the authors' productivity, quality, career length, and proportion of self-citations on the authors' h -index?

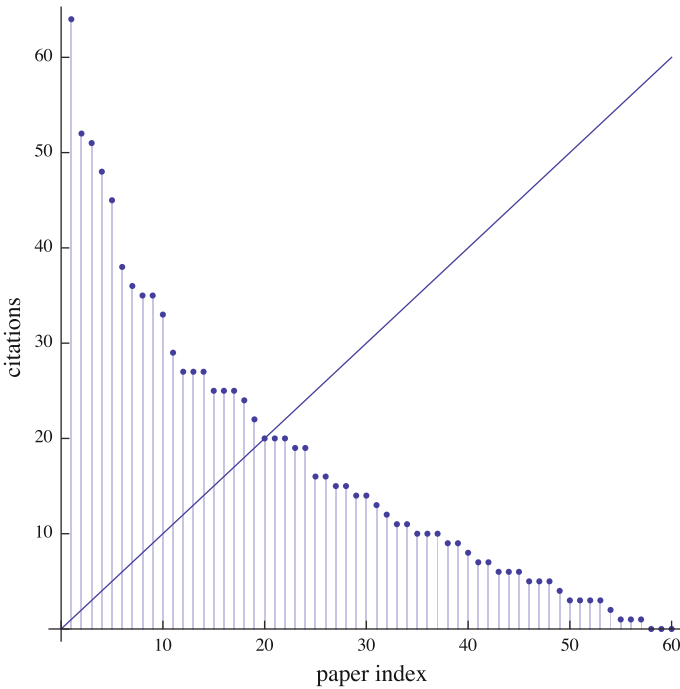


Fig. 1 Citation profile of an example author

Method

To be able to investigate how far the *h*-index can be inflated we need to consider extreme authors that focus all their self-citations on increasing their *h*-indexes. We are currently not aware of a sufficient number of such extreme authors to be able to appropriately answer all our research questions with data from real authors. The only exception might be Ike Antkare, an mock scientist who is only citing himself [10] We therefore did no base our analyses on existing authors, but focused on simulated authors.

Wolfgang Glänzel and his co-authors [6, 7] proposed a stochastic model for the publishing and citation process. However, here we make use of a more recent stochastic model proposed by Burrell [3], which is better suited for our simulation. The main result of this model as described in Eq. 1 defines is the expected number of papers that receive at least *n* citations by time *T*:

$$E(n, T) = \theta T \left(1 - \frac{\alpha}{(v-1)T} \sum_{r=0}^{n-1} B\left(\frac{T}{\alpha+T}, r+1, v-1\right) \right) \quad \text{for } n = 1, 2, 3, \dots, \quad (1)$$

with $B(x; a, b)$ the regularized incomplete beta function defined as

$$B(x; a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^x y^{a-1} (1-y)^{b-1} dy, \quad (2)$$

and $\Gamma(a)$ the gamma function. The model depends on quality parameters which characterize a certain author:

- T is the time passed since the start of the researchers career
- θ is the productivity (mean number of publications per unit time)
- v is the standard shape parameter of the citation distribution (gamma distribution), which is related to its height.
- $1/\alpha$ is the standard scale parameter of the citation distribution (gamma distribution), which is related to its width.
- $\frac{v}{\alpha}$ is the mean citation rate (average number of citations per paper per year)
- n is the number of citations

Equation 1 can be considered as the average citedness of papers from authors of a given quality (see Fig. 2). The graph has the expected shape for citation indexes and appears to match reality. It follows the well documented skew [16] and hence we assume face validity of the model. We invert the expression in Eq. 1 to create a theoretical citation profile for an example author characterized by these quality parameters.

We used Burrell's model to simulate the publication process of an average author defined through the parameters mentioned above. We added one parameter μ , which is defined as the number of self-citations per paper. For practical reasons, we defined μ as a constant, but it is conceivable that the number of self-citations might change over the duration of a scientist's career. We implemented three different self-citation strategies:

- (1) the author makes μ strategic self citations by the method described above (unfair condition)
- (2) the author cites his μ last papers (fair condition)
- (3) the author randomly cites μ of papers (random condition)

The fair condition is based on the observation that the number of self-citations is the highest for new papers and declines rapidly over time [1]. The random condition provides

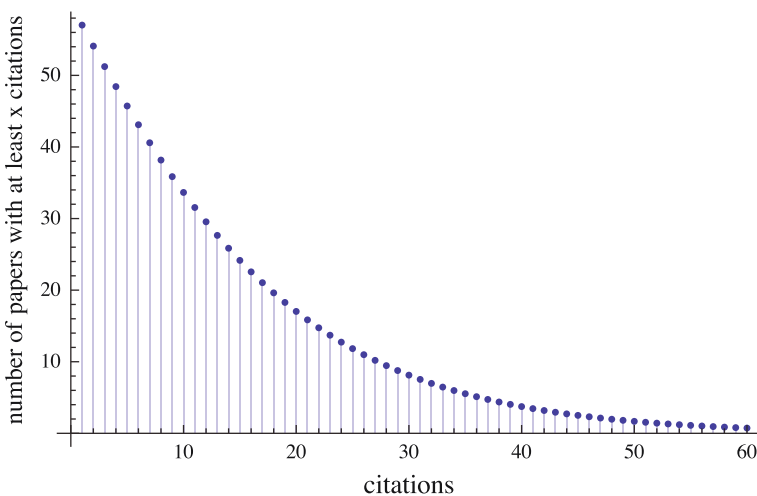


Fig. 2 Average citedness of papers from authors of productivity $\theta = 3$, career length $T = 20$, mean citation rate $\frac{v}{\alpha} = \frac{3}{2}$ with $v = 3$ and $\alpha = 2$

us with a baseline against which we can compare the other two conditions. The simulation, implemented in Mathematica, consists of a main loop that cycles for the p published papers from 1 to $\theta \times T$ through the following steps:

- (1) calculate the current h_p -index of the author
- (2) calculate the citations received from other researchers through Burrell’s model
- (3) place μ self citations through one of the three strategies described above
- (4) calculate indicators, such as the q -index described below, for the current state
- (5) sum the citations from others and the self-citations

Results

To answer the question how much an authors can inflate their h -index through self-citations we first would like to present an archetypical author. He publishes three papers per year over a total of 20 years and he makes three self-citations per paper. Figure 3 shows how the h -index develops over the period of publishing each of the 60 papers. After 20 years the author would have an h -index of 19 if he had used the unfair strategy, while a random self-citation strategy would have resulted in an h -index of only 14. Through the strategic placement of his self-citations, he was able to inflate his h -index by 5.

If we now look at the citation index of the unfair author, we notice a humpback around the h -paper, which is in this case the 19th paper (see Fig. 4). An author with a random self-citation strategy does not have such a humpback. This may come at no surprise, since the humpback is a direct result of self-citing papers close to the h -paper.

To be able to assess the size of the humpback we propose the q -index. Quasimodo, a fictional character in Victor Hugo’s novel “The Hunchback of Notre Dame”, inspired its name. Quasimodo has a severely hunched back, which reminded us of the humpback in the citation profile. In comparison to the penalty system proposed by Burrell [2] the q -index

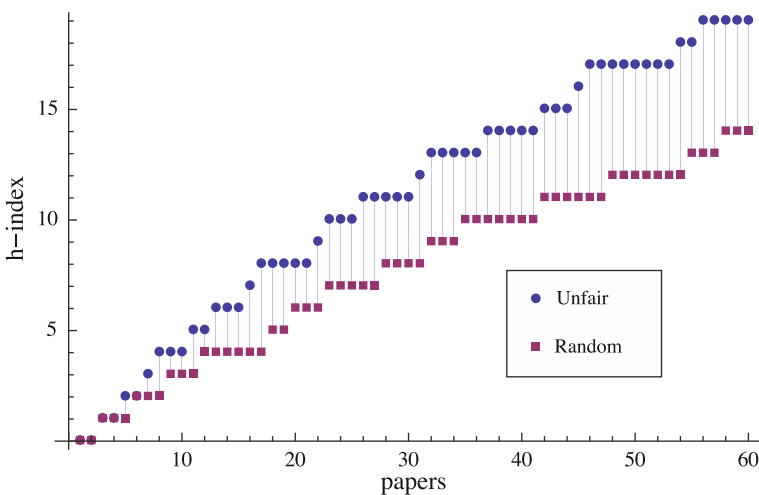


Fig. 3 Development of h_p -index over published papers p for an author with $\theta = 3$, career length $T = 20$, mean citation rate $\frac{\nu}{z} = \frac{3}{2}$ with $\nu = 3$ and $\alpha = 2$

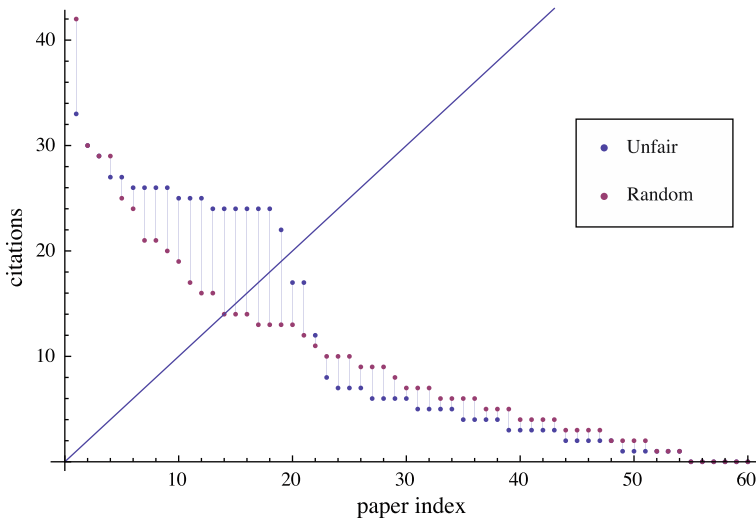


Fig. 4 Citation profile $c_{60,i}$ over paper index i of an author in the unfair and in the random condition with $\theta = 3$, career length $T = 20$, mean citation rate $\frac{\mu}{\alpha} = \frac{3}{2}$ with $\nu = 3$ and $\alpha = 2$, and for a total number of published papers of $p = \theta \times T = 60$

does not decrease the citation count, but it introduces a stand alone indicator for the self-citation behavior.

The q -index can be calculated as follows. First, sort all papers ($i = 1 \dots p$) of an author or organization, given a certain number of already published papers p , according to their citations in a descending order: $c_{p,i}$. This creates the well known citation profiles, as shown in Fig. 1. This citation profile is characterized by h_p . For each self-citation of a paper that has equal or fewer citations than the h_p -paper, the author receives a q -score. This q -score is calculated by dividing 1 by the number of different citations scores between the h_p -paper and the paper that receives the self-citation. If the author cites the h_p -paper(s) then the score will be $\frac{1}{1}$. If he cites paper(s) that have the next fewer citations, then he receives a score of $\frac{1}{2}$ and so forth. Next papers i which have the same citation score $c_{p,i}$ as the previous one, receive the same q -score. The formal definition is given by:

$$q_{p,i} = \begin{cases} 0 & i < h_p \\ \frac{1}{i+1-a_{p,i}-h_p} & i \geq h_p, \end{cases} \tag{3}$$

with $a_{p,i}$ given by

$$a_{p,i} = \begin{cases} 0 & i \leq h_p \\ a_{p,i-1} & i > h_p, c_{p,i} - c_{p,i-1} \neq 0 \\ a_{p,i-1} + 1 & i > h_p, c_{p,i} - c_{p,i-1} = 0. \end{cases} \tag{4}$$

Note that we only take into account the q -scores for the actually cited papers i , and therefore the summed q -score that an author receives for publishing a new paper p can only range between 0 and μ .

Lets take an example to illustrate the q -scores. Figure 5 shows the citation profile of our archetypical unfair author. The x axis lists the q -scores that this author receives for citing his own papers. Notice that the author does not receive any q -score for self-citing papers

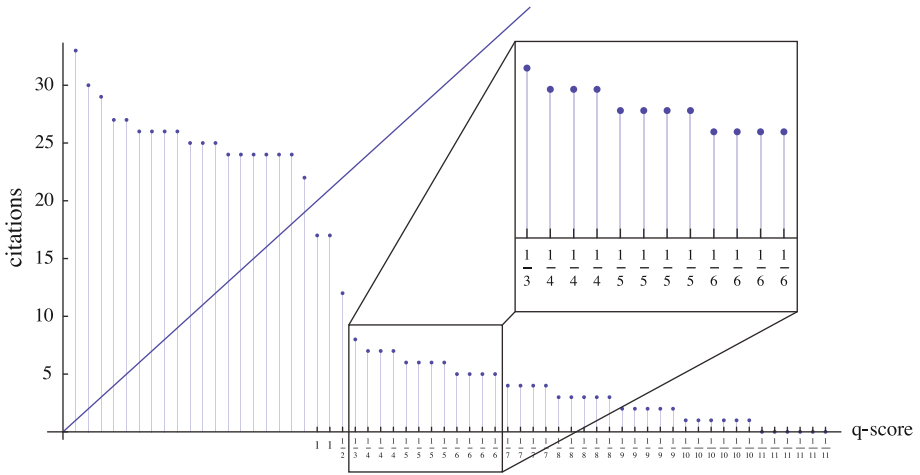


Fig. 5 Unfair citation profile of Fig. 4 with the q -scores on the x axis

that have more citations than the h_p -paper. These papers are on the left of the diagonal h -line. Citing these papers does not directly inflate the h -index and are therefore not considered when calculating q -scores. Also notice that papers that have the same number of citations also receive the same q -scores. Their order can be assumed to be random and hence it would not be fair to give them different q -scores.

We plotted the q -scores in the order in which the papers were published (see Fig. 6). If the author publishes a new paper that cites three of his own papers, then the three q -scores he received are summed. The paper index on the x axis thereby defines the order in which the papers were published. Initially, all three self-citing strategies produce the same q -scores. This comes at no surprise since the fourth published paper can only cite its three predecessors. Only starting from the fifth paper, the author can choose which paper not to cite. A few papers later, we find significant differences between the three self-citation conditions. The unfair author receives high q -scores with very little spread, since he is always citing very close to the h_p -paper.

The author with a fair self-citing strategy receives lower and lower q -scores (see Fig. 6). This can be explained by the fact that the total number of publications q grows much faster

Fig. 6 Summed q -score indexes over published paper p , for the unfair, fair and random condition

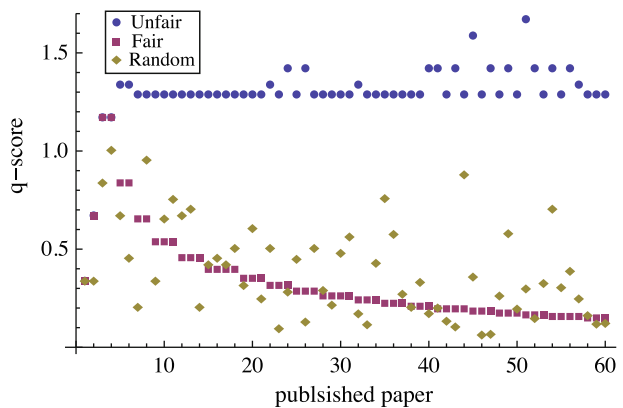
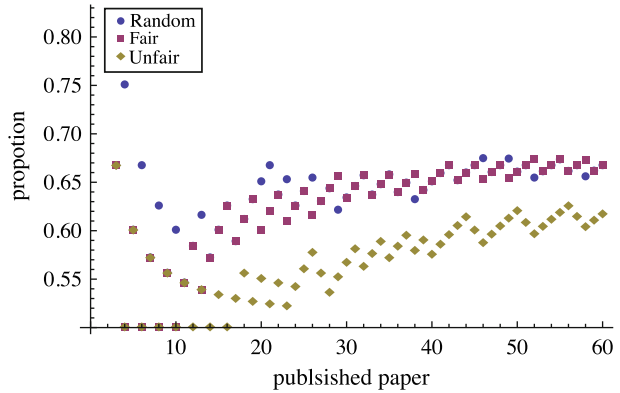


Fig. 7 Proportion of papers with fewer citations than the h_p -paper



than the h -index. The proportion of papers that have fewer citations than the h_p -paper (to the right of the h_p -paper) to the papers that have equal or more citations than the h_p -paper (from the h_p -paper to the left) is increasing (see Fig. 7). The new papers that the fair author cites become further and further away from the h_p -paper and hence attract lower and lower q -scores.

An author with a random self-citation strategy has a much higher spread in his q -scores, but they also appear to decrease. The growing number of papers that have fewer citations than the h_p -paper can also explain this trend. The papers in this long tail cause lower and lower q -scores (see Fig. 7).

We propose the q -index as the summed q -scores the author received for each self-citation s ranging from 0 to the total number of self-citations μ , in published paper j , to a paper in the citation profile indexed by $i_{j,s}$. This is normalized by the number of published papers p :

$$Q_p = \frac{1}{p} \sum_{j=1}^p \sum_{s=0}^{\mu} q_{j,i_{j,s}} \tag{5}$$

The normalization by p assures that the q -index is approximately constant over all published papers if an author consistently cites according to the unfair scheme. This linear behavior can be seen from the unnormalized q -index in Fig. 8 for the unfair condition, while in the fair and the random condition it flattens out and are in general far below the unnormalized q -index of the unfair condition (see Fig. 8). Interestingly, the curve for the fair and the random condition are very close to each other. It might be difficult to distinguish between authors that use these two strategies. The q -index's range follows as:

$$\leq Q_p \leq \mu \tag{6}$$

The q -index should be accompanied by the standard deviation of the summed q -scores. For our example of our archetypical author, the q -indexes at $p = 60$ are available in Table 1. The q -index of the fair and random condition are within one standard deviation from each other. We may therefore conclude that the q -index is not able to detect a significant differences between these two conditions. The q -index for the unfair condition is approximately ten standard deviations away from the q -index of the random condition and approximately four standard deviations away from the fair condition. It would be very unlikely if the difference observed would be due to chance. To test this hypothesis, we performed the non-parametric Mann–Whitney test, since we cannot assume a normal

Fig. 8 Unnormalized *q*-index $p \times Q_p$ over published papers *p*, for the unfair, fair and random condition

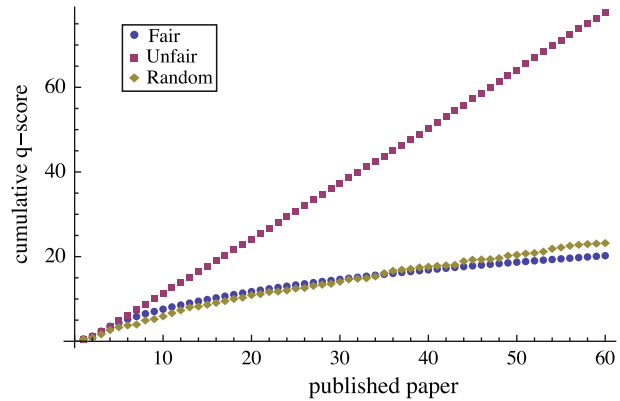


Table 1 *q*-Index and standard deviation across all conditions

Condition	<i>Q</i>	Std dev.
Random	0.475	0.076
Fair	0.509	0.149
Unfair	1.176	0.179

distribution of the data. The distributions in the random and unfair conditions differed significantly (Mann–Whitney $U = 41.5$, $n_1 = n_2 = 60$, $P < 0.01$, two-tailed).

Next, we were interested in how the different parameters of Burrell’s model influence the development of the *h*-index. We started by varying the productivity θ from one paper per year to eighteen papers per year. These values seem plausible minimum and maximum values. Of course, an director of a research institute that insists on co-authorship of every paper produced in his/her institute may exceed these boundary conditions, but the analysis of honorary authorship are not in the focus of our study. The other parameters remained at their stereotypical setting of career length $T = 20$, mean citation rate $\frac{\nu}{\alpha} = \frac{3}{2}$ with $\nu = 3$ and $\alpha = 2$. Figure 9 shows that *h*-index quickly increases $0 < \theta < 5$ and then slowly flattens. An author that publishes six papers per year will have an more than double the *h*-index compare to an author that publishes only one paper per year. The unfair strategy benefits in particular by an increased productivity, since more published papers also mean more self-citations.

The next parameter we varied is the career length T between 1 and 40 years, which again seemed plausible boundary conditions. The remaining parameters were set to the stereotypical values of $\theta = 3$, mean citation rate $\frac{\nu}{\alpha} = \frac{3}{2}$ with $\nu = 3$ and $\alpha = 2$. Figure 10 shows a linear increase for the *h*-index for all three conditions. The *h*-score increases by approximately one per year.

We varied the number of self-citations per papers μ from one to ten, which appeared to be reasonable limits. The other parameters remained at their stereotypical settings. The results displayed in Fig. 11 show that μ has a smaller effect on the *h*-index compared to θ and T . In the fair and random condition, the increasing μ results on only a mild increase in the *h*-index. In the unfair condition, the *h*-index grows over μ , but again less compared to θ and T . The small effect size is also visible in absolute terms. With ten self-citations per paper, an unfair author is only able to get up to an *h*-index of around 30, while he can get up to 50 with a publication rate of 18 papers per year.

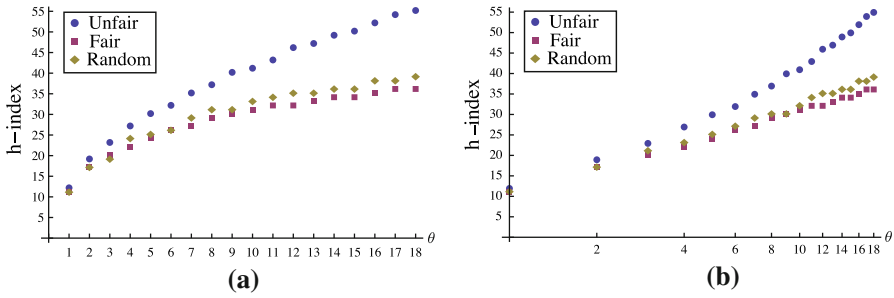


Fig. 9 **a** h -Index across the productivity θ . **b** Same, but on a logarithm productivity scale. On this scale, the fair and random citation strategies confirm the straight lines as also observed by Burrell. The unfair strategy, however, clearly deviates from the linear behavior

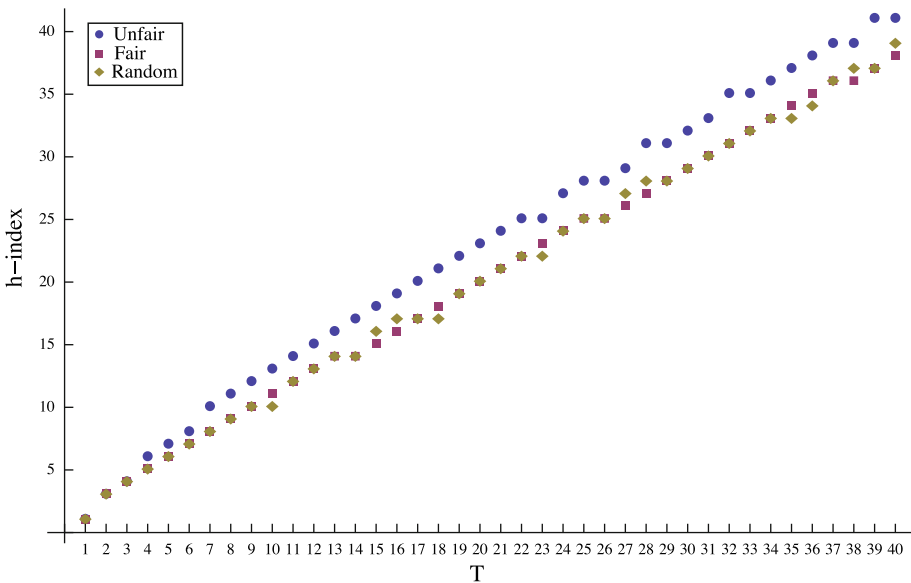


Fig. 10 h -Index across the career length T

Next, we changed the shape parameter ν and the scale parameter α of the citation distribution, keeping in mind that $\frac{\nu}{\alpha}$ is the mean citation rate, which defines how many citations a paper receives from other researchers. We increased the value for ν and α from one to ten, which appeared reasonable boundary conditions. The other parameters remained at their stereotypical settings. Figure 12 shows that the increasing value for ν increases the number of citations from others, which in turn negates the advantage of strategic self-citations. At ν , there is no more difference between the unfair condition and the other two conditions. For authors that produce highly esteemed works by others, strategic self-citations have little positive effect. Burrell offered a similar result in his Fig. 4(a) he kept α at 5 and increased ν from 5 to 500.

When increasing the value for α , the mean citations rate drops, which has the opposite effect from increasing ν . And indeed, Fig. 13 shows that the h -index decreases over an

Fig. 11 *h*-Index across the number of self citation μ

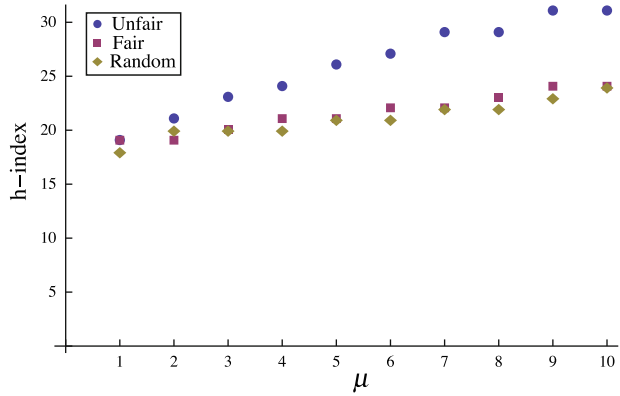


Fig. 12 *h*-Index across the height of the citation distribution parameter ν

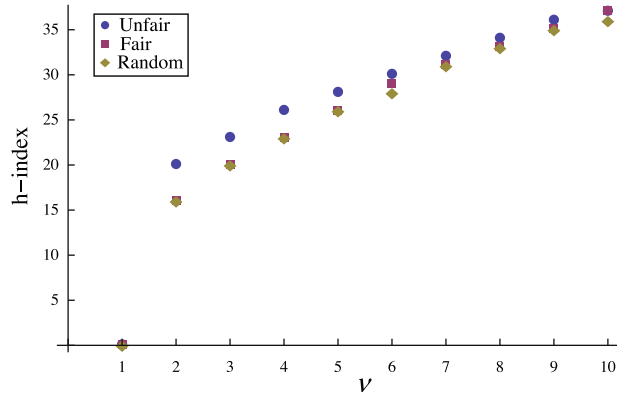
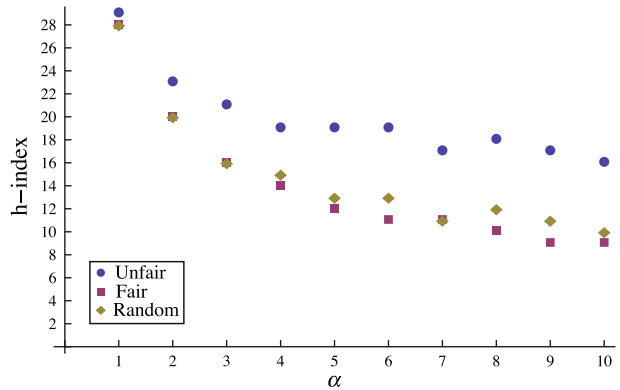


Fig. 13 *h*-Index and *q*-scores across the width of the citation distribution parameter α



increasing value for α . The gap between the unfair condition and the other two conditions increases, indicating that making strategic self-citations becomes increasingly beneficial.

To assess how strong the effect of productivity, career length, number of self-citations, and mean citation rate, is on the *h*-index, we calculated the average change in the *h*-index as:

Table 2 The average Δ and standard deviation for all parameters

Parameter	Condition	Mean Δ	Std dev. Δ
θ	Unfair	2.53	1.50
	Fair	1.47	1.42
	Random	1.65	1.69
T	Unfair	1.03	0.58
	Fair	0.95	0.32
	Random	0.97	0.67
μ	Unfair	1.33	0.87
	Fair	0.56	0.53
	Random	0.44	0.53
ν	Unfair	4.11	5.99
	Fair	4.11	4.51
	Random	4.00	4.58
α	Unfair	-1.44	2.01
	Fair	-2.11	2.52
	Random	-2.00	2.65

$$\Delta_k = h_k - h_{k-1}, \quad (7)$$

where h_k is the the h -index when the parameter $(\theta, T, \mu, \nu, \alpha)$ is k , ranging from 2 to the maximum of the respective parameter. The average Δ_k and its standard deviation is displayed in Table 2. The mean citation rate has the strongest impact on the h -index. The increase of ν by one increases the h -index on average by four and an increase in α of one decreases the h -index by around two. The second strongest effect stems from the productivity of the author. By publishing one paper more per year, the author's h -index increases by approximately 1.5. With every year passed, the h -index increases on average by one. The number of self-citations has only a strong effect for authors that strategically place them. For all other authors, it has the smallest benefit.

Conclusions

The results of our simulation show that authors can significantly inflate their h -index, and possible also other indices, by strategically citing their own publications. Calculating the q -index helps identifying such behavior and plotting the individual q -scores over the sequence of published papers allows us to gain additional insights into the publication history of an author. The q -index also allows us to run standard statistical test for cases that are ambiguous. The unfair author in our study is an extreme example and real authors might apply more subtle strategies to manipulate their h -index. The q -index also conveniently ranges from 1 to μ , which gives it an easy to interpret range. Our simulation is able to provide the benchmark of a random self-citation behavior, which can be used to compare the real authors' q -index against.

Overall we can conclude that the unfair self-citation strategy is mainly useful for authors that are less productive and that attract less citations from others. The most effective method to increase one's h -index is to produce work that is highly cited by other.

The next best strategy is to be productive. The length of the career only has a moderate influence. On average, authors can increase their *h*-index by one per year, as it was predicted by Hirsch [9].

This study does have some limitations. We have to acknowledge that our simulations has not yet been verified by comparing its results to data from real authors. However, we hope to test the *q*-index on real authors in the next phase of our project.

The application of the *q*-index to real data may prove to be difficult, since it requires knowledge of each publication, including the date of each received citations. While this is relatively easy to accomplish in a simulation, the data from the real world has a tendency to be incomplete and occasionally ambiguous. A first application could be achieved on the well structured data from the Web of Science and in a second phase, attempts could be made to parse the results from Google Scholar.

We also have to consider that the mean citation rate does not model the size of the research community in which a certain author may operate. This potentially influential factor is not part of Burrell's model and hence we are unable to make any judgements about it. We are also not able to make any judgements about the differences between scientific disciplines.

In essence, we showed that the *h*-index is vulnerable to manipulations by self-citations. We propose the *q*-index as a metric to judge how strategic the self-citations of an author have been. In addition, we showed that the best way to increase one's *h*-index is to write interesting papers. This might be no surprise, but sometimes it is necessary to even state the obvious.

Open Access This article is distributed under the terms of the Creative Commons Attribution Noncommercial License which permits any noncommercial use, distribution, and reproduction in any medium, provided the original author(s) and source are credited.

References

1. Aksnes, D. (2003). A macro study of self-citation. *Scientometrics*, 56(2), 235–246.
2. Burrell, Q. L. (2007). Should the *h*-index be discounted? *ISSI Newsletter*, 3-S, 65–67.
3. Burrell, Q. L. (2007). Hirsch's *h*-index: A stochastic model. *Journal of Informetrics*, 1(1), 16–25.
4. Egghe, L. (2006) Theory and practise of the *g*-index. *Scientometrics*, 69(1), 131–152 doi:[10.1007/s11192-006-0144-7](https://doi.org/10.1007/s11192-006-0144-7).
5. Garcia-Perez, M. (2009). A multidimensional extension to Hirschs *h*-index. *Scientometrics*, 81(3): 779–785.
6. Glänzel, W., & Schoepflin, U. (1994). A stochastic model for the ageing of scientific literature. *Scientometrics*, 30(1), 49–64.
7. Glänzel, W., & Schubert, A. (1995). Predictive aspects of a stochastic model for citation processes. *Information Processing and Management*, 31(1), 69–80.
8. Glänzel, W., Thijs, B., & Schlemmer, B. (2004). A bibliometric approach to the role of author self-citations in scientific communication. *Scientometrics*, 59(1), 63–77.
9. Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, 102(46), 16569–16572.
10. Labbe, C. (2010). Ike Antkare one of the great stars in the scientific firmament. *ISSI Newsletter*, 6(2), 48–52.
11. Meho, L. I., & Yang, K. (2007). Impact of data sources on citation counts and rankings of lis faculty: Web of science versus scopus and google scholar. *Journal of the American Society for Information Science and Technology* 58(13), 2105–2125. doi:[10.1002/asi.20677](https://doi.org/10.1002/asi.20677).
12. Van Noorden, R. (2010). Metrics: A profusion of measures. *Nature*, 465, 864–866.
13. Panaretos, J., & Malesios, C. (2009). Assessing scientific research performance and impact with single indices. *Scientometrics*, 81(3), 635–670.
14. Prathap, G. (2010). Is there a place for a mock *h*-index? *Scientometrics*, 84(1), 153–165.

15. Schreiber, M. (2007). Self-citation corrections for the Hirsch index. *Europhysics Letters*, 78(3). 0002.
16. Seglen, P. O. (1992). The skewness of science. *Journal of the American Society for Information Science*, 43(9), 628–638.
17. Silverman, B. W. (2009). Comment: Bibliometrics in the context of the UK research assessment exercise. *Statistical Science* 24(1), 15–16.