Insights Into the Relationship Between the *h*-Index and Self-Citations

Ernesto Gianoli
Departamento de Botánica, Universidad de Concepción, Casilla 160-C, Concepción, Chile. Center for Advanced Studies in Ecology & Biodiversity (CASEB), P. Universidad Católica de Chile, Santiago, Chile. E-mail: egianoli@udec.cl.

Marco A. Molina-Montenegro
Departamento de Botánica, Universidad de Concepción, Casilla 160-C, Concepción, Chile. E-mail: marcmoli@udec.cl

We analyze the publication output of 119 Chilean ecologists and find strong evidence that self-citations significantly affect the *h*-index increase. Furthermore, we show that the relationship between the increase in the *h*-index and the proportion of self-citations differs between high and low *h*-index researchers. In particular, our results show that it is in the low *h*-index group where self-citations cause the greater impact.

The h-index (Hirsch, 2005) was proposed to quantify the research productivity of individual scientists and is defined as follows: “A scientist has index h if h of his/her N_p papers have at least h citations each, and the other $(N_p - h)$ papers have no more than h citations each” (p. 16569). The h-index is increasingly recognized as a simple yet sound estimator of the scientific research output of individuals, but there is also disagreement on its reliability (Kelly & Jennions, 2006; Bornmann & Daniel, 2007; Hirsch, 2007; Lehmann Jackson, & Lautrup, 2006). There is evidence that the h-index is sensitive to inflation through self-citations (Kelly & Jennions; Schreiber, 2007a,b; Enqvist & Frommen, 2008). The proponent of the h-index originally acknowledged this point but underrated its importance. (“While self-citations can obviously increase a scientist’s h, their effect on h is much smaller than on the total citation count.” (Hirsch, 2005, p. 16571)). A recent study (Engqvist & Frommen, 2008) analyzed the publication record of 40 researchers and argued that the supposed sensitivity of the h-index is overestimated and that the purported increase of h-index via self-citations is rather unlikely. Here, we first show that Engqvist and Frommen’s analysis is partly misleading and that their conclusions are biased. Second, we present an alternative, more meaningful analysis based on a broader database, which confirms that self-citations significantly affect the h-index increase. Furthermore, we show that the relationship between the increase of h-index and the proportion of self-citations differs between high and low h-index researchers. This should be taken into account when using the h-index for academic committee decisions on fellowships, appointments, and promotions.

Engqvist and Frommen (2008) randomly selected 40 authors from the fields of ecology and evolution and identified the citation causing their last increase in h-index. Then they found the first citation appearing afterwards that would have caused the same increase in h-index. Their rationale was that the time elapsed between the appearance of the two citations reflects the time that the h-index is dependent on one single citation and, hence, estimates the duration of the effect of selective self-citation of a given paper. They found that half of all h-increasing citations were “redundant” within 2 months. This analysis is flawed. First, it does not discriminate between self-citations and non-self-citations. Given that such “effect duration” most likely varies depending on the origin of the citations, no clear conclusion can be drawn. Thus, if both the first and the second citations are self-citations, then the time elapsed almost certainly would be shorter and, hence, the effect duration would be, in average, of lower magnitude than in the case of the two non-self-citations.

Received November 7, 2008; revised December 8, 2008; accepted January 9, 2009

© 2009 ASIS&T • Published online 13 February 2009 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/asi.21042

The advantages in terms of h-index increase for this hypothetical self-citer would vary with time; but, if this is a consistent citation pattern, then, overall, it would certainly confer advantages over a less self-referential author. Second, even demonstrating that such inflation of h-index due to self-citations vanishes with time does not challenge the main point underlying this discussion: that a reliable indicator of scientific productivity, upon which decisions for fellowships and appointments may be taken, must be free from self-citation biases. The problem is not to prove that a selective citation strategy directed towards increasing h-index is not always successful. The point is to be aware of the effect of self-citation behaviors on a fair estimation of scientific impact for comparative purposes. Anyway, Engqvist and Frommen's analysis shows a highly significant effect of the average number of self-citations per paper on the inflation of h-index due to self-citations. However, they interpreted it as "rather modest," without any quantitative basis, despite the fact that an increase from two to four self-citations per paper would make an increase of two in the h-index. This may be significant for committee decisions on fellowships or appointments if other estimators of the competing researchers are similar.

To evaluate the link between self-citation behavior and the h-index, we carried out two analyses. We first regressed the increase in h-index during a 6-month period against the proportion of total citations that are self-citations. Two categories of h-index, low and high, were considered and also included in the analysis to determine whether the increase in h-index was more likely in high h-index researchers. We further tested the significance of the relationship between h-index increase and self-citations separately for high and low h-index researchers. To build up the data set we included the publication record of 119 Chilean ecologists, which were selected from membership lists of Chilean societies of ecological sciences as well as from exhaustive surveys in Web sites of academic departments of ecology, biology, botany, and zoology. We only considered active researchers (i.e., with at least three papers published in the last 3 years) and set at 10 the minimum number of published papers for a researcher to be entered in the database. Consequently, it was not a random sample; we attempted to include all ecologists *sensu lato* complying with those requisites. We took great care in getting the exact publication list of a given scientist by both avoiding the inclusion of papers by homonymous individuals and including papers published under different combinations of surnames or initials. Data were compiled from the Web of Science (Thomson ISI) in the third week of February 2008 and in the third week of August 2008. To calculate the proportion of total citations that are self-citations we used the "Citation Report" option from the Web of Science. For each author, we first retrieved the sum of the times cited (T) and then recorded the number of cites that are not self-citations (E) by choosing the "view without self-citations" option. The proportion of self-citations was calculated as $[1 - (E/T)]$. This parameter takes into account self-citations only by the author under analysis. Thus, it does not consider citations made by coauthors, and, hence, it is a conservative estimation of self-citations (see Schreiber 2007a). The final data set included researchers with a mean h-index of 9.2 ± 0.5 s.e. (range: 2–29). The mean h-index was used as a reference point to cluster the groups of ecologists in low- h-index (from 2 to 8) and high h-index (from 9 to 29).

The h-index increase recorded after 6 months for 119 Chilean ecologists was between 0 and 3 (mean ± s.e.: 0.55 ± 0.06) and the proportion of self-citations ranged from 0.07 to 0.70 (mean ± s.e.: $0.33 \geq 0.01$). Using a generalized linear model (Poisson distribution linked to a log function) we confirmed that there was a greater increase in h-index for researchers with a greater proportion of self-citations (estimate: $3.01 \geq 0.87$, Wald statistic $= 11.95$, $df = 1$, $p < 0.001$; proportion data were arc-sin transformed prior to analysis). In the same analysis, the category of h-index, low (2–8) or high (9–29), did not affect h-index increase (Wald statistic $= 0.12$, $df = 1$, $p > 0.72$). Thus, authors with greater h-indices do not have a greater probability of increasing them. We found it more interesting that the significant relationship between increase in h-index and proportion of self-citations was largely due to the low h-index Chilean ecologists. Thus, whereas a greater proportion of self-citations resulted in a greater increase of h-index (estimate: 2.89 ± 1.01, Wald statistic $= 8.21$, $df = 1$, $p < 0.005$) for researchers with low h-index (2–8, $n = 70$), there was no relationship between h-index increase and proportion of self-citations (estimate: 2.65 ± 2.21, Wald statistic $= 1.44$, $df = 1$, $p > 0.22$) for researchers with high h-index (9–29, $n = 49$). In the low h-index group, researchers with null increase of h-index had a median proportion of self-citations of 0.25 (range: 0.07–0.60), those with an increase of 1 had a median of 0.33 (range: 0.14–0.70), those with an increase of 2 had a median proportion of self-citations of 0.32 (range: 0.23–0.43), and the only ecologist showing an increase of h-index of three had a proportion of self-citations of 0.44. A mechanistic explanatory analysis of this pattern is out of the scope of this contribution, but it might be related to the structure of the algorithm or distribution underlying h-index calculation (Hirsch, 2005, 2007). We find it interesting that after addressing self-citation corrections for the h-index in some small data sets, Schreiber (2007b) speculated that for young scientists with comparatively low h-index, the influence of self-citations should be relatively strong. We now provide empirical support to his guess.

In conclusion, it is clear that differences in self-citation behavior make a difference in the h-index outcome, as has been reported earlier (Kelly & Jennions, 2006; Schreiber, 2007a,b; Enqvist & Frommen, 2008). We think that the main issue is not to quantify the sensitivity of the h-index to self-citations, because criteria to interpret the outcome of such analysis may be subjective. The major point is that the effect of self-citations must be taken into account for a fair estimation of the impact of a researcher's publications. Thus, we endorse the contention that a sharpened h-index not considering self-citations should be preferred (Schreiber, 2007a). This is particularly important when the h-index is included in the evaluation criteria for faculty appointment and promotion. Furthermore, our results show that it is in the low *h*-index group where self-citations cause the greater impact. This is the group where most applicants for a first academic job are included. It follows that an enhanced index should replace the *h*-index, at least in the case of academic appointment processes.

References

Bornmann, L., & Daniel, H-D. (2007). What do we know about the *h* index? Journal of the American Society for Information Science and Technology, 58(9), 1381–1385.

Engqvist, L., & Frommen, J.G. (2008). The *h*-index and self-citations. Trends in Ecology and Evolution, 23(5), 250–252.

Hirsch, J.E. (2005). An index to quantify an individual's scientific research output. Proceedings of National Academy of Sciences, USA, 102(46), 16569–16572.

Hirsch, J.E. (2007). Does the *h* index have predictive power? Proceedings of National Academy of Sciences, USA, 104, 19193–19198.

Kelly, C.D., & Jennions, M.D. (2006). The *h* index and career assessment by numbers. Trends in Ecology and Evolution, 21(4), 167–170.

Lehmann, S., Jackson, A.D., & Lautrup, B.E. (2006). Measures for measures. Nature, 444, 1003–1004.

Schreiber, M. (2007a). A case study of the Hirsch index for 26 non-prominent physicists. Annalen der Physik, 16, 640–652.

Schreiber, M. (2007b). Self-citation corrections for the Hirsch index. EPL, 78, 30002-p1-30002-p6.

DOI: 10.1002/asi