Review of: "Measuring researchers' success more fairly: going beyond the H-index"

Robert Sternberg[1]

1 Cornell University

It is admirable that Dr. Formoso wishes to propose options beyond the h-index proposed by Hirsch (2005). The h-index is useful in some respects and flawed in other respects, as Dr. Formoso recognizes and as many others have recognized as well (e.g., Bornmann et al., 2008; Enago Academy, 2022; Gingras & Khelfaoui, 2020; Ioannidis et al., 2016; Koltun & Hafner, 2021; Kreiner, 2016; Sternberg, 2016, 2017; Waltman, 2016). On the one hand, the h-index is flawed in certain ways. But on the other hand, as the author points out, *all* quantitative indices are flawed in some way. There is no perfect index of scholarly impact or of anything else.

Some of the flaws that critics have identified are, at the very least, arguable. For example, although it is true that the h-index is highly correlated with author productivity, so is the level of scientific and other creative contributions (Simonton, 1977, 1984, 1985, 1988, 1997). Simonton has shown a strong relation between quantity and quality in many domains, so it is not clear exactly what the "flaw" is of the h-index.

Another criticism is the effect of self-citations. One could correct for self-citations; but the problem is that, in small fields, much of the work may have been done by relatively few people, including the scholar whose h-index is being evaluated. So, one would also have to control for the number of people working in a field, an area of a field, a sub-area of a field, and so on. And even within a field, a scholar may be unique in using a novel approach or making a novel argument, so that the scholar may need to cite themselves to make whatever point they may wish to make.

Dr. Formoso points out that one could correct for position of authorship in particular articles. This is fair enough, as some authors may achieve a higher h-index by making small contributions to large numbers of publications. And some scientists are rarely first authors. But again, any such correction would be fraught. Methodologists often make valuable contributions to papers but usually are not first authors on substantive papers where they handled much of the methodology or data analysis. In some papers, the most important author is the last author, but in others, the last author is the least important author. It often is hard to tell. One senior scholar may defer to junior scholars and give them higher positions of authorship, figuring they need it more for their careers than the senior scholar does; but another senior scholar may be egocentric and always want to be first author. In "big science," there may be hundreds of authors. It is really hard to tell, on average, what position of authorship means.

I could go through each argument against the h-index, but my general points are that (a) there is no perfect index; (b) some of the criticisms may be justified but it is not clear what to do about them; (c) other criticisms are dubious; and (d)

the h-index probably has lasted as long as it has for lack of a replacement that is easily understandable, computable, and usable.

One possible route to managing the limitations of the h-index is to point out its limitations, or to suggest an imperfect replacement. Another is to realize, as Formoso points out, that all indices are flawed, and that in correcting the flaws of one index, one may end up introducing a new index that differs primarily in that it is flawed in other ways than the h-index.

We also sometimes forget why the h-index was welcomed in the first place. In the past, scientists sometimes had relatively little idea of how much scholarly impact a scholar's work was having. In a tenure meeting, an unproductive or low-impact scholar who felt threatened by a scholar with higher productivity or impact would and sometimes did make the case that the higher-impact scholar published "too much," or was not doing "high-quality" work, or whatever. The main protection the h-index has offered is against the rampant jealousy and envy in science, where colleagues sometimes try to bring each other down to protect their own challenged self-esteem. The h-index does not provide total protection against mediocre scholars trying to bring down better ones, but at least it provides some measure of who is accomplishing what. Yet, all is not well with the use of the h-index or any other single index.

I believe a useful analogy for the h-index is to IQ. There are countless articles and books on the limitations of IQ, some of which I have written myself (e.g., Sternberg, 1997). But IQ still is used because it has been found to predict various criteria and thus to be useful (Sackett et al., 2020). There have been various attempts to replace IQ (and its proxies, such as SAT and ACT), but so far, none has quite caught on. An alternative approach, which follows from Formoso's own arguments, is not to try to replace IQ but rather to recognize its limitations and the fact that any single index will be limited, no matter what it is or how it is computed. The question then becomes what else ones want to know about a person. For example, although many selective colleges and universities use standardized tests, including IQ proxy tests, in admissions, they also look at many other aspects of performance, such as grade-point average (GPA), teacher and guidance counselor letters of recommendation, extracurricular activities, personal essays, interviews, and the like.

I believe that a better use of scholarly time than pointing out the problems of h (they have been written about many times) might be to find constructive supplements for it, much as colleges and universities have done for standardized test scores. Any single index, no matter what it is, will be limited and flawed. By using converging operations (Garner et al., 1956) and a multi-trait, multi-method approach (Campbell & Fiske, 1959) to evaluate the scientific contribution of an individual, one will gain a better sense of what the individual has to offer. There already are lots of citation indices—e.g., h-index, g-index, i10, total number of citations). It is not clear we need more of them, as they all measure roughly the same thing, just as different tests of general mental ability measure roughly the same thing, regardless of their cosmetic differences. If one wishes to have a better sense of impact, the best thing to do is to use multiple measures, such as h, g, i10, and total # of citations. No one index will be perfect, but a combination of indices will be better than any one single index.

I do not believe the future of better assessment of scholarly contributions is in yet some other modification of the h-index, which in the end is likely to be highly correlated with the h-index (much as proxy measures like the SAT and ACT are highly correlated with IQ—Frey & Detterman, 2004; Koenig et al., 2008). The various indices will be highly correlated

because they are also measuring the same basic construct of impact through scientific citations. Rather, it is to decide what it is we value in scientists, and to start with measures of those constructs—creativity in generating novel and useful ideas, problem-finding in recognizing important problems, wisdom in applying the ideas, analytical rigor of a scientist's work, collegiality, service to an institution or field, teaching skills, or whatever. We will do a better job in measuring scientific or other contributions not through yet another variant of the h-index, but rather in deciding what matters and then figuring out the best ways to assess it. Oddly, before the advent of the h-index, tenure and promotion meetings looked at many of these variables. Often such meetings still do look at diverse criteria, but it is too easy for committee members or others to get sucked into believing that a single index can tell us more than it really can tell us. I have referred to this as the quantitative precision fallacy (Sternberg, 1997).

In sum, we can do better than the h-index. But the solution to doing better is not in creating endless variants of the h-index. The solution rather is in deciding what we care about in evaluating a scientist, and then finding the best ways to assess whatever it is we care about. One of the things we care about may be impact as measured by various citation indices, including h. But there is more, and our responsibility is to identify what the "more" is and then optimize on our ways of measuring the many qualities besides indexed impact that we care about.

References

- Bornmann, L., Mutz, R., Daniel, H.D. (2008). Are there better indices for evaluation purposes than the h index? A comparison of nine different variants of the h index using data from biomedicine. *Journal of the American Society for Information Science and Technology*, 59(5), 830–837. https://doi.org/10.1002/asi.20806

- Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*(2), 81–105. https://doi.org/10.1037/h0046016

- Enago Academy (2022, November 3). H-index: Is it enough to score scientific excellence? https://www.enago.com/academy/drawbacks-of-h-index/

- Frey, M. C., & Detterman, D. K. (2004). Scholastic assessment or *g*? The relationship between the Scholastic Assessment Test and general cognitive ability. *Psychological Science, 15,* 373-378.

- Garner, W. R., Hake, H. W., & Eriksen, C. W. (1956). Operationism and the concept of perception. *Psychological Review*, *63*, 149–159.

- Gingras, Y., & Khelfaoui, M. (2020, July 8). Why the h-index is a bogus measure of scientific impact. *The Conversation,* https://theconversation.com/why-the-h-index-is-a-bogus-measure-of-academic-impact-141684

- Hirsch J. E. (2005). An index to quantify an individual's scientific research output. Proc. Natl. Acad. Sci. U.S.A. 102, 16569–16572. 10.1073/pnas.0507655102

- Ioannidis, J.P.A., Klavans, R., Boyack, K.W. (2016). Multiple Citation Indicators and Their Composite across Scientific Disciplines. *PLoS Biology* 14(7): e1002501 https://doi.org/10.1371/journal.pbio.1002501

- Koenig, K.A., Frey, M.C. & Detterman, D.K. (2008). ACT and general cognitive ability. *Intelligence 36,* 153-160.

- Koltun, V., Hafner, D. (2021) The h-index is no longer an effective correlate of scientific reputation. *PLoS ONE* 16(6):

e0253397. https://doi.org/10.1371/journal.pone.0253397

- Kreiner, G. (2016, November 2). *Frontiers in Human Neuroscience,* 10: 556, doi: 10.3389/fnhum.2016.00556

- Sackett, P. R., Shewach, O. R., & Dahlke, J. A. (2020). The predictive value of general intelligence. In R. J. Sternberg (Ed.), *Human intelligence: An introduction* (pp. 381-414). Cambridge University Press.

- Simonton D. K. (1977). Creative productivity, age, and stress: a biographical time-series analysis of 10 classical composers. *J. Pers. Soc. Psychol* 35 791–804. 10.1037/0022-3514.35.11.791

- Simonton D. K. (1984). Creative productivity and age: a mathematical model based on a two-step cognitive process. *Dev. Rev.* 4 77–111. 10.1016/0273-2297(84)90020-0

- Simonton D. K. (1985). Quality, quantity, and age: The careers of ten distinguished psychologists. *Int. J. Aging Hum. Dev.* 21 241–254. 10.2190/KB7E-A45M-X8X7-DQJ4

- Simonton D. K. (1988). Age and outstanding achievement: what do we know after a century of research? *Psychol. Bull.* 104 251–267. 10.1037/0033-2909.104.2.251

- Simonton D. K. (1997). Creative productivity: a predictive and explanatory model of career trajectories and landmarks. *Psychol. Rev.* 104 66–89. 10.1037/0033-295X.104.1.66 Waltman L. (2016). A review of the literature on citation impact indicators. *Journal of Informetrics,* 10(2), 365-391. https://doi.org/10.1016/j.joi.2016.02.007

- Sternberg, R. J. (1997). *Successful intelligence.* Plume.

- Sternberg, R. J. (2016). "Am I famous yet?" Judging scholarly merit in psychological science: An introduction. *Perspectives on Psychological Science, 11,* 877-881*.*

- Sternberg, R. J. (2017). Afterword: In the matter of judging scholarly merit. *Perspectives on Psychological Science, 12, https://doi/full/10.1177/1745691617720729*.