Article available at http://www.europhysicsnews.org or http://dx.doi.org/10.1051/epn/2009704

* **Franck Laloë** [1] and **Remy Mosseri** [2],
* [1] Laboratoire Kastler Brossel, ENS, CNRS and UPMC, Paris, France
* [2] Laboratoire de Physique Théorique de la Matière Condensée, UPMC and CNRS, Paris, France
* DOI: 10.1051/epn/2009704

BIBLIOMETRIC EVALUATION OF INDIVIDUAL RESEARCHERS: NOT EVEN RIGHT… NOT EVEN WRONG!

Scientific research is probably, among the various human activities, one of those that are most subject to benchmarking and evaluation (refereeing of articles, research contracts, evaluation of laboratories, etc.). In recent years, an increasing weight has been given to bibliometry, yielding various rankings and numbers. While such data may sometimes bring useful information, in case of evaluation of individuals, sadly the implementation often seems to arise from a loss of critical and rational mind.

Nobody would suggest that the long-term artistic impact of contemporary writers, composers, painters, etc. should be assessed by using quantitative methods (number of books sold in bookshops, tickets in concerts, number of paintings, etc.). One would in this way get some information on the quality of the artistic production, but also mostly detect other characteristics of the productions, as well as fashions; the most creative artists, those who will remain famous for decades or centuries, would probably not emerge. Neither would anyone suggest evaluating quantitatively the efficiency of state officers by the number of administrative reports they write!

Although scientific research is neither art nor administration, quantitative methods of evaluation of individual scientists suffer from the same problem: they contain some information, but with only a relatively small part that corresponds to scientific quality. The use of indices such as the H-index [1] at the level of individuals is easy and therefore attractive, but mostly unscientific. An even more serious problem is that the generalization of quantitative evaluation of research will create (and actually is already creating) perverse feedback effects, decreasing the quality of publications and therefore affecting the general functioning of research.

Wolfgang Pauli

It is reported that Wolfgang Pauli, one of the geniuses who created quantum mechanics, was very irritated after reading an uninteresting article and exclaimed: *"it is not right, and not even wrong!"*. Verification and falsification are indeed at the heart of Nature sciences. Pauli's remark could apply as well to many applications of bibliometry, proposed by those who believe that, as soon as numbers are manipulated, they create scientific reasoning.

The bibliometric evaluation of researchers is "not even wrong": yes, if we compare an internationally well-known scientist to an eccentric who has never been cited by anyone but himself, the bibliometric indices of the former will be much better than for the latter; no doubt about it. Therefore, if the purpose was to distinguish between exceptional from mediocre scientists, it would undoubtedly be possible to use bibliometric methods. One would then find again… what was known before. But, if the idea is to obtain really useful information, for instance to rank researchers within a relatively homogeneous group (members of a laboratory for instance), then the situation is different: one immediately notices the existence of large fluctuations of the indices [1,2] (H, G, etc.) that are surprising. Significantly different values are obtained by scientists whose production is judged similar by the scientific community. Why?

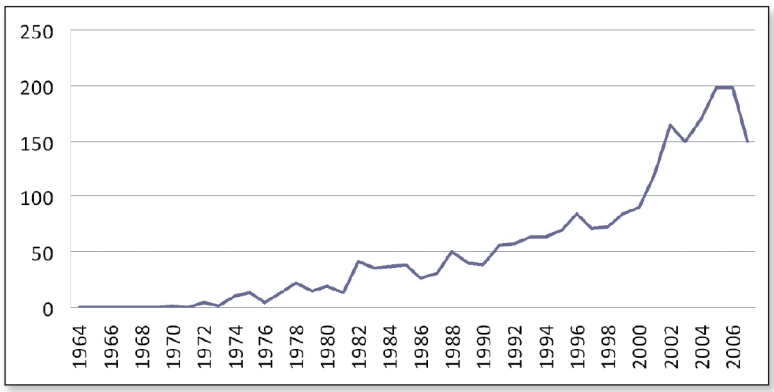
Extracting signal from noise

Several reasons explain why bibliometric methods provide a simplistic view of individual scientific contributions. They do contain information about scientific quality, but this "signal" is buried in a "noise" created by a dependence on many other variables. Let us take for instance the H-index, which is a function of a first variable X that we assume to correspond more or less to the scientific quality, of a variable Y related to the personal style of the person (working in collaboration or not, preference for "fashionable" scientific subjects, interest in applications, etc.), of a variable Z related to the publication style (preference for short letters or long, more developed, articles; for prestigious journals of general interest such as Nature or Science, even if they are not much used in the domain), and of a variable W

(does the person belong to a prestigious school of research, or has preferred to try and create a new field of research?). Obviously, this list of variables is not limitative; one could for instance add the taste of the person for scientific congress, which is not always connected with creativity. Any scientist knows that the only way to determine the target variable X from H is to eliminate the noise by averages. It they are performed over a large sample of persons, variables Y, Z, … will take all possible values and their influence will average out, leaving the target variable X to emerge. This is the reason why bibliometric methods allow one to obtain relevant data at large scales, for instance the comparison between two big fields of research in different countries. Nevertheless, using H to determine X with a single statistical sample is merely a scientific error that would not be accepted in any serious laboratory.

Real purpose of references

Moreover, this evaluation method relies on an implicit postulate that, actually, is by no way obvious. The idea is that the references contained in all scientific articles should be some sort of prize list among all published articles in the relevant scientific domain; this ranking could then be used as measure of quality averaged over the opinion of many authors. But this is certainly not the way authors create the list of references in their articles: including them is a scientific act, which has little to do with bibliometry or ranking. The real purpose is to give the reader the information that is needed to understand the article. This is also a strongly contextual process: for instance, an article will be cited by convenience, because it allows a shortening of the article under writing - review articles will then be preferred to the original sources. Similar articles are also preferred, independently of their scientific interest, because they simplify the writing… and avoid tensions with colleagues. Authors may even cite articles that they consider as wrong, because they wish to correct their errors. In experimental disciplines, articles describing methods and apparatuses are also favoured. By contrast, big scientific discoveries, abstract new ideas, are rarely cited through the big original article, but rather through daughter publications inspired by the first. It is therefore a very indirect use of citations, probably even a complete misunderstanding of their function, to take them as the only basic element for evaluating scientific quality. What is even worse, it creates perverse side effects since this method of evaluation tends to react back on the way scientific citations are given in articles, in a way that does not improve the quality of scientific communication.



▼ **FIG. 1:** Number or citations per year of the famous article by John Bell "On the Einstein-Podolsky-Rosen paradox" (Physics, vol. 1, 195-200 (1964) until mid 2007. While this article was almost not cited for many years, 30 years later it became more and more cited. In particular, this article would have not contributed at all to the IF (impact factor) of the journal in which it was published!

Various bias

In "hard" sciences, the data base that is used for bibliometry is mostly the SCI [3]. A first problem immediately comes to mind: books are not taken into account in the calculation of the H factor that this basis provides in 2 clicks! This is very strange, since most scientists agree that one of the best ways for a creative researcher to influence a scientific domain is precisely to publish books.

A second problem: the indices G, H, etc. that are usually used to rank individuals are as biased as the well-known Shanghai ranking. In these indices, the contribution of an author is exactly the same whether he/she is the only author, or if he/she has 10 co-authors! Calculating indices G', H', etc… from the number of citations divided by the number of authors may look elementary and even more logical, but no one seems to be doing the calculation in this way with the SCI [4]. The biasing problem is then obvious: if three friends decide to put together all their publications during their entire career, all their H-index will be strongly increased.

Third problem: the prominent weight of short term in bibliographic indices. In many scientific domains, minor technical breakthroughs create a flurry of publications, which may sometimes become quickly forgotten. As a consequence, the bibliometric indices are very sensitive to fashions. This problem could also easily be cured: it would be easy to calculate indices that favour articles that have a long term influence, for instance by including only articles that are still cited after 3 of 5 years; but no one seems to be doing it. A striking illustration of the long time constants involved in the citation rate of very influential articles in physics is given in Figure 1.

Fourth problem, rather technical but nevertheless real: the SCI data basis is not homogeneous, since the entry of data has fluctuated in time with the persons in charge of it. It therefore requires a specialist to make the necessary series of corrections, which we cannot discuss here, but who cares: it is so much easier to get a ranking with three mouse clicks!

In addition to being of mediocre quality, this superficial use of bibliometry creates more serious problems at a deeper level than just biased evaluation. It may affect the quality of scientific communication between researchers, and therefore in the long run the quality of research itself. It is clear that, if the criteria for selection favour the short term bibliometric impact, most scientists will adapt to this fact and orient their work in a direction that will provide good indices, even at the

expense of scientific quality. Since the process of scientific communication is directly at the heart of the functioning of science, it will shift the general effort towards short term impact, reducing the progress of knowledge in the long run. This is actually already happening: some scientific journals, in order to improve their "Impact factor", implement a biased citation policy, by letting the authors know that "if your manuscript contains at least 4 citations of articles in our journal, it will be more easily accepted". What can be worse for scientific quality and clarity?

Conclusions

This leads us to the following conclusions about these indices for evaluation of an individual:

1. They have not been rationally tested with sufficient care, in comparison with other methods of evaluation. Paradoxically, the methods for evaluating research have escaped the scientific selection that is applied in all disciplines! Even elementary consistency checks as those mentioned above (e.g. effect of dividing by the number of authors) have not been performed; no one really understand the influence of the arbitrary method of calculation on the final result.
2. No one seems to have had the time to honestly try to obtain indices that are more closely related to the quality and long-term relevance of science.
3. They are "not even wrong" since they undoubtedly do contain some information on individuals, mostly when this information is trivial and already known. When the indices are used in real life in a homogeneous population, they give more information on other variables than the quality of the work, such as the style of work of the evaluated person.
4. Their success is not related to the quality of the information they provide, but more on the facility: a saving of the time necessary to make a real evaluation.
5. Finally, it seems that the faith in these indices is not very different from something that escapes rationality. One could compare this with astrology and numerology, which pretend to be scientific but have never gone through the process of scientific selection.

A recipe

If you are a bad researcher, there is probably no way to get a good H-index, so this is not for you. If you are a good researcher, and if you want a better H-index, here is some advice:

1. Choose to work in a group of at least 5 or 6 colleagues, if possible more, who publish all their articles in common; this should allow you to significantly increase your index (you will all have the same index, but who cares?). Moreover, this will allow you to share material and human benefits (postdocs, for instance) which may also increase your real productivity. No need to mention that, the more prominent these colleagues are, the more you will improve your H!
2. Favour big scientific domains over smaller ones; since small domains cite more large domains than the converse, a positive correlation exists between the citation rate of articles and the size of the scientific domain. Above all, avoid choosing subjects that are not mainstream; even if you are a genius, it will take at least 10 years to see your work appreciated. Moreover, your articles will not be cited more than those which your article has stimulated. In short, do not take too much scientific risk!
3. Above all, do not waste your time in publishing books, whatever their international intellectual impact is; they play no role in the usual H factor.
4. Do not give too much importance to what was initially your motivation to do science, namely the production of new and original knowledge, in particular when you write articles; keep in mind that public relations are more important. ■

Big scientific discoveries are rarely cited through the **original article**

About the Authors

Franck Laloë is an emeritus "directeur de recherches" at Laboratoire Kastler Brossel, ENS (Paris). He has done research on optical pumping, spin polarized quantum gases, nuclear spin waves, the calculation of the effects of interactions on the Bose-Einstein transition temperature, and recently on quantum non-local effects in Bose-Einstein condensates.

Rémy Mosseri is "directeur de recherches" at laboratoire de physique de la matière condensée, CNRS-UPMC Paris. He worked on disordered systems, quasicrystals, and is now interested in quantum information. He also published a biography of the physicist Léon Brillouin.

The first version of this article was published in *Reflets de la Physique* **13**, 23 (2009)

notes

[1] The H-index attempts to measure both the scientific productivity and the apparent scientific impact of a scientist, a group, a laboratory, etc. J.E. Hirsch introduced it in 2005 and defined it in the following way: *"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 at most h citations each"*. The "Web of Science" now gives direct access to the H index in a few mouse clicks.

[2] The G index was introduced in 2006 by L. Egge, as an alternative to the H index. Its definition is *"Given a set of articles ranked in decreasing order of the number of citations that they received, the g-index is the (unique) largest number such that the top g articles received **on average** at least g citations"*.

[3] SCI is for "Science Citation Index", the data base on which the commercial service of the WOS (Web of Science) is offered by ISI./Thomson.

[4] Our point is not to imply that this method would solve all problems, and even that it would be sufficient. The indices G', H', ... would not necessarily be more relevant than G,H,…; what is interesting is that they would change the results of rankings, illustrating the arbitrary character of the final result.