Bibliometrics

Quality Versus Quantity: Assessing Individual Research Performance

José-Alain Sahel[1,2,3,4,5,6,7]*

[1]INSERM, U968, Paris F-75012, France.

[2]Université Pierre et Marie Curie-Paris 6, UM80, Institut de la Vision, Paris F-75012, France.

[3]CNRS, UMR 7210, 75012 Paris, France.

[4]Centre Hospitalier National d'Ophtalmologie des Quinze-Vingts, INSERM-DHOS CIC 503, Paris F-75012, France.

[5]Fondation Ophtalmologique Adolphe de Rothschild, 75019 Paris, France.

[6]Institute of Ophthalmology, University College of London, London EC1V 9EL, UK.

[7]Académie des Sciences–Institut de France, 75006 Paris, France.
*Corresponding author. E-mail: j.sahel@gmail.com.

Bibliometrics could potentially improve objective evaluation of individual researchers, if qualitative measurements are included.

ABSTRACT

Evaluating individual research performance is a complex task that ideally examines productivity, scientific impact, and research quality—a task that metrics alone have been unable to achieve. In January 2011, the French Academy of Sciences published a report on current bibliometric (citation metric) methods for evaluating individual researchers, as well as recommendations for the integration of quality assessment. Here, we draw on key issues raised by this report and comment on the suggestions for improving existing research evaluation practices.

BALANCING QUANTITY AND QUALITY

Evaluating individual scientific performance is an essential component of research assessment, and outcomes of such evaluations can play a key role in institutional research strategies, including funding schemes, hiring, firing, and promotions. However, there is little consensus and no internationally accepted standards by which to measure scientific performance objectively. Thus, the evaluation of individual researchers remains a notoriously difficult process with no standard solution. Marcus Tullius Cicero once wrote, "Non enim numero haec iudicantur, sed pondere" (*1*). Translation: The number does not matter, the quality does. In line with Cicero's outlook on quality versus quantity, the French Academy of Sciences analyzed current bibliometric (citation metric) methods for evaluating individual researchers and made recommendations in January 2011 for the integration of quality assessment (*2*). The essence of the report is discussed in this Commentary.

Evaluation by experts in the field has been the primary means of assessing a researcher's performance, although it can be biased by subjective factors, such as conflicts of interest, disciplinary or local favoritism, insufficient competence in the research area, or superficial examination. To ensure objective evaluation by experts, a quantitative analytical tool known as bibliometry (science metrics or citation metrics) has been integrated gradually into evaluation processes (Fig. 1). Bibliometry started with the idea of an impact factor, which was first mentioned in *Science* in 1955 (*3*), and has evolved to weigh several aspects of published work, including journal impact factor, total number of citations, average number of citations per paper, average number of citations per author, average number of citations per year, the number of authors per paper, Hirsch's *h*-index, Egghe's *g*-index, and the contemporary *h*-index. The development of science metrics has accelerated recently, with the availability of online databases used to calculate bibliometric indicators, such as the Thomson Reuters Web of Science (http://thomsonreuters.com/), Scopus (http://www.scopus.com/home.url), and Google Scholar (http://scholar.google.com/). Within the past decade, metrics have secured a foothold in the evaluation of individual, team, and institutional research because the use of such metrics appears to be easier and faster than the qualitative assessment by experts. Because of the ease of use of various metrics, however, bibliometry tends to be applied in excessive and even incorrect ways, especially when used as standalone analyses.

The French Academy of Sciences (FAS) is concerned that some of the current evaluation practices—in particular, the uncritical use of publication metrics—might be inadequate for evaluating individual scientific performance. In its recent review (*2*), the FAS addressed the advantages and limitations of the main existing quantitative indicators, stressed that judging the quality of a scientific work in terms of conceptual and technological innovation of the research is essential, and reaffirmed its position about the decisive role that experts must play in research assessment (*2*, *4*). It also strongly recommended that additional criteria be taken into consideration when assessing individual research performance. These criteria include teaching, mentoring, participation in collective tasks, and collaboration-building, in addition to quantitative parameters that are not measured by bibliometrics, such as number of patents, speaker invitations, international contracts, distinctions, and technology transfers. It appears that the best course of action will be a balanced combination of the qualitative (experts) and the quantitative (bibliometrics).

BIBLIOMETRICS: INDICATORS OR NOT?

Bibliometrics use mathematical and statistical methods to measure scientific output; thus, they provide a quantitative—not a qualitative—assessment of individual research performance. The most commonly used bibliometric indicators, as well as their strengths and weaknesses, are described below.

Impact factor

The impact factor, a major quantitative indicator of the quality and popularity of a journal, is defined by the median number of citations for a given period of the articles published in a journal. The impact factor of a journal is calculated by dividing the number of current-year citations by the source items published during the previous two years (*5*). According to the FAS, the impact factor of journals in which a researcher has published is a useful but highly controversial indicator of individual performance (*2*). The most common issue is variation among subject areas; in general, a basic science journal will have a higher average impact factor than journals in specialized or applied areas. Individual article quality within a journal is also not reflected by a journal's impact factor because citations for an individual paper can be much higher or lower than what might be expected on the basis of that journal's impact factor (*2*, *6*, *7*). In addition, self-citations are not corrected for when calculating the impact factor (*6*). On account of these limitations, the FAS considers the tendency of certain researchers to organize their work and publication policy according to the journal in which they intend to publish their article to be a dangerous practice. In extreme situations, such journal-centric behavior can trigger scientific misconduct. The FAS notes that there has been an increase in the practice of using journal impact factors for the evaluation of an individual researcher for the purpose of career advancement in some European countries, such as France, and in certain disciplines, such as biology and medicine (*2*).

Number of citations

The number of times an author has been cited is an important bibliometric indicator; however, it is a value that has several important limitations. First, citation number depends on the quality of the database used. Second, it does not consider where the author is located in the author list. Third, sometimes articles can have a considerable number of citations for reasons that might not relate to the quality or importance of the scientific content. Fourth, articles published in prestigious journals are privileged as compared with those with equal quality but published in journals of average notoriety. Fifth, depending on cultural issues, advantage can be given to citations of scientists from the same country, to scientists from

other countries (in particular Americans, as often is the case in France), or to articles written in English rather than in French, for example (*2*). For these cultural reasons, novel and important papers might attract little attention for several years after their publication. Lastly, citation numbers also tend to be greater for review articles than for original research articles. Self-citations do not reflect the impact of a publication and should therefore not be included in a citation analysis when this is intended to give an assessment of the scientific achievement of a scientist (*8*).

*New indicators (*h*-index,* g*-index)*

Recently, new bibliometric indicators borne out of databases indexing articles and their citations were introduced to address the needs of objectively evaluating individual researchers. In 2005, Jorge Hirsch proposed the *h*-index as a tool for quantifying the scientific impact of an individual researcher (*9*). The *h*-index of a scientist is the number of papers co-authored by the researcher with at least *h* citations each; for example, an *h*-index of 20 means that an individual researcher has co-authored 20 papers that have each been cited at least 20 times each. This index has the major advantage to measure simultaneously the scientist's productivity (number of papers published over years) with the cumulative impact of the scientist's output (the number of citations for each paper). Although the *h*-index is preferable to other standard single-number criteria (such as the total number of papers, total number of citations, or number of citations per paper), it has several disadvantages. First, it varies with scientific fields. As an example, *h*-indices in the life sciences are much higher than in physics (*9*). Second, it favors senior researchers by never decreasing with age, even if an individual discontinues scientific research (*10*). Third, citation databases provide different *h*-indexes as a result of differences in coverage, even when generated for the same author at the same time (*11*, *12*). Fourth, the *h*-index does not consider the context of the citations (such as negative findings or retracted works). Fifth, it is strongly affected by the total number of papers, which may underestimate scientists with short careers and scientists who have published only a few although notable papers. The *h*-index also integrates every publication of an individual researcher, regardless of his or her role in authorship, and does not distinguish articles of pathbreaking or influential scientific impact. Contemporary *h*-index (referred to as *hc*-index), as suggested by Sidiropoulos *et al*. (*10*), takes into account the age of each article and weights recently published work more heavily. As such, the *hc*-index may offer a fairer comparison between junior and senior academics than the regular *h*-index (*13*).

The *g*-index was introduced (*14*) to distinguish quality, giving more weight to highly cited articles. The *g*-index of a scientist is the highest number *g* of articles (a set of articles ordered in terms of decreasing citation counts) that together received g^2 or more citations; for example, a *g*-index of 20 means that 20 publications of a researcher have a total number of citations of at least 400. Egghe pointed out that the *g*-index value will always be higher than the *h*-index value, making easier to differentiate the performance of authors. If Researcher A has published 10 articles, and each has received 4 citations, the researcher's *h*-index is 4. If the Researcher B has also written 10 articles, and 9 of them have received 4 citations each, the researcher's *h*-index is also 4, regardless of how many citations the 10th article has received. However, if the tenth article has received 20 citations the *g*-index of the Researcher B would be 6; for 50 citations, the *g*-index would be 9 (*15*). Thus, one or several highly cited articles can influence the final *g*-index of an individual researcher, thus highlighting the impact of authors.

CHOOSING AN INDICATOR

Bibliometry is easy to use because of its simple calculations. However, it is important to realize that the purely bibliometric approaches are inadequate because no indicator alone can summarize the quality of the scientific performance of a researcher. The use of a set of metrics (such as number of citations, h-index, or g-index) would give a more accurate estimation of the researcher's scientific impact. At the same time, metrics should not be made too complex because they can become a source of conceptual errors that are then difficult to identify. FAS discourages the use of metrics as a standalone evaluation tool, the use of only one bibliometric indicator, the use of the journal's impact factor to evaluate the quality of an article, neglecting the impact of the scientific field/sub-field, and ignoring author placement in the case of multiple authorship (*2*).

In 2004, INSERM (the French National Institute of Health and Medical Research) introduced bibliometrics as part of its research assessment procedures. Bibliometric analysis is based on publication indicators that are validated by the evaluated researchers and are at the disposal of the evaluation committees. In addition to the basic indicators (citation numbers and journal impact factor), the measures used by INSERM include the number of publications in the first 10% of journals ranked by decreasing impact factor in a given field (top 10% impact factor, according to Thomson Reuters Journal Citation Reports) and the number of publications from an individual researcher that fall within the top 10% of articles (ranked by the total citations) in annual cohorts from each of the 22 disciplines defined by Thomson Reuters Essential Science Indicators. All indicators take into account the research fields, the year of publication, and the author's position among the signers by assigning an index of 1 to the first or last author, an index of 0.5 for the second or the next to last author, and 0.25 for all other author positions. Notably, the author's index can only be used in biomedical research because for other fields the rank of the authors may follow different rules, such as in physics, in which they are listed in alphabetical order.

Bibliometric indicator interpretation requires competent expert knowledge of metrics, and in order to ensure good practice, INSERM trains members of evaluation committees on state-of-the-art science metric methods. INSERM has noted that correlation analysis of publication—in other words, scoring by members of evaluation committees and the use of any bibliometric indicator alone—is rather low. For example, the articles of all teams received a number of citations irrespective of the journal in which they were published, with only low correlation between the journal impact factor and the number of times each publication was cited. No correlation was found between the journal impact factor and the individual publication citations, or between the "Top 1%" publications and the impact factor (*16*). INSERM analysis emphasizes the fact that each indicator has its advantages and limitations, and care must be taken not to consider them alone as "surrogate" markers of team performance. Several indicators must be taken into account when evaluating the overall output of a research team. The use of bibliometric indicators requires great vigilance; but, according to the INSERM experience, metrics enrich the evaluation committees' debates about the scientific quality of team performance (*16*).

As reported by the FAS, bibliometric practices vary considerably from country to country. A worldwide *Nature* survey (*17*) emphasized that 70% of the interviewed academic scientists, department heads, and other administrators believe that bibliometrics are used for recruitment and promotions, and 63% of them consider the use of these measures to be inadequate. Many Anglo-Saxon countries use bibliometrics to evaluate performances of universities and research organizations, whereas for hiring and promotions, the curriculum vitae, interview process, and letters of recommendation "count" more than the bibliometric

indicators (*2*). In contrast, metrics are used for recruiting in Chinese and Asian universities in general, although movement toward the use of letters of recommendation is currently underway (*2*). In France, an extensive use of publication metrics for individual and institutional evaluations has been noted in the biomedical sciences (*2*).

Research evaluation practices also vary by field and subfield owing in part to the large disparities across community sizes and the literature coverage provided by citation databases. As reviewed by the FAS, evaluation of individual researchers in the mechanical sciences, computing, and applied mathematics fields includes both the quality and the number of publications, as well as scientific awards and the number of invitations to speak at conferences, software, patents, and technology transfer agreements. Organization of scientific meetings and editorial responsibilities are also taken into consideration. The younger researchers are evaluated by experts during interviews and while they give seminars. In these fields, publication does not always play a leading role in transferring knowledge; thus, during a long professional career, metrics give rather weak and inaccurate estimation of research performance. Bibliometrics are therefore used only as a decision-making aid, but not as a main tool for evaluation.

In physics and its subfields, evaluation methods vary. In general, a combination of quantitative (number of publications, *h*-index) and qualitative measures (keynote and invited speeches, mentoring programs) plays a decisive role in the evaluation of senior scientists only. In astrophysics, metrics are largely used for evaluation, recruiting, promotions, and funding allocations. In chemistry, the main bibliometric indicators (*h*-index, total number of citations, and number of citations per article) are taken into consideration when discussing the careers of senior researchers (those with more than 10 to 12 years of research activity). In recruiting young researchers, experts interview the candidate to examine ability to present and discuss the subject matter proficiently; the individual's publication record is also considered. However, the national committees for chemistry of French scientific and university institutions [Centre National de la Recherche Scientifique (CNRS) and Conseil National des Universités (CNU), respectively] usually avoid bibliometrics altogether for an individual's evaluation.

In economics, evaluation by experts in the field plays the most important role for recruitments and promotions, but bibliometric indicators are used to help this decision-making. For the humanities and social sciences (philosophy, history, law, sociology, psychology, languages, political sciences, and art) and for mathematics, the existing databases do not cover these fields sufficiently. As a consequence, these fields are not able to properly use bibliometrics. In contrast, in biology and medicine the quantitative indicators—in particular the journal's impact factor—are widely used for evaluating individual researchers (*2*).

STRATEGIES AND RECOMMENDATIONS

The FAS acknowledged that bibliometrics could be a very useful evaluation tool when handled by experts in the field. According to its recommendations, the use of bibliometrics by monodisciplinary juries should be of nondecisive value; instead, the experts of these evaluation committees know the candidates well enough to compare more precisely and objectively the individual performance of each of them. In the case of pluridisciplinary (interdisciplinary) juries, bibliometrics can be successfully used, but only if the experts consider the differences between scientific fields and subfields (as mentioned above). For this purpose, the choice of indicators and the methodology to evaluate the full spectrum of research activity of a scientist should be initially validated. As emphasized by the FAS, bibliometrics should not be used for deciding which young scientists to recruit. In addition,

the bibliometric set should be chosen depending on the purpose of the evaluation: recruitment, promotion, funding allocation, or distinction. Calculations should not be left to nonspecialists (such as administrators that could use the rapidly accessible data in a biased way) because the number of potential errors in judgement and assessment is too large. Frequent errors to be avoided include the homonyms, variations in the use of name initials, and the use of incomplete databases. It is important that the complete list of publications be checked by the researcher concerned. Researchers could even be asked to produce their own indicators (if provided with appropriate guidelines for calculation); these calculations should subsequently be approved. The evaluation process must be transparent and replicable, with clearly defined targets, context, and purpose of the assessment.

To improve the use of bibliometrics, a consensus has been reached by the FAS (*2*) to perform a series of studies and to evaluate various methodological approaches, including (i) retrospective studies to compare decisions made by experts and evaluating committees, with results potentially obtained by bibliometrics; (ii) studies to refine the existing indicators and bibliometric standards; (iii) authorship clarification; (iv) development of standards for originality and innovation; (v) discussion on the citation discrepancies on the basis of geographical- or field-localism; (vi) monitoring the bibliometric indicators of outstanding researchers (a category reserved for those who have made important and lasting research contributions their specific field and who have obtained international recognition); (vii) examining the prospective values of the indicators for researchers that changed their field orientation with time; (viii) examining the indicators of researchers receiving great awards such as Nobel Prize, Fields Medal, and medals of notorious academies and institutions; (ix) studies on how bibliometrics affect the scientific behavior of the researchers; and (x) establishment of standards of good practice in the use of bibliometrics for analyzing individual research performance.

FIXING THE FLAWS

Assessing research performance is important for recognizing productivity, innovation, and novelty and plays a major role in academic appointment and promotion. However, the means of assessment—namely, bibliometrics—are often flawed. Bibliometrics have enormous potential to assist the qualitative evaluation of individual researchers; however, none of the bibliometric indicators alone (or even considering a set of them) allow for an acceptable and well-balanced evaluation of the activity of a researcher. The use of bibliometrics should continue to evolve through in-depth discussion on what the metrics mean and how they can be best interpreted by experts in the given scientific field.

Fig. 1. Can individual research performance be summarized by numbers?
CREDIT: IMAGE COURTESY OF D. FRANGOV (FRANGOV DIMITAR PLAMENOV COMPANY)

References and Notes

<bok>1. M. T. Cicero, *De Officiis*. Translated by Walter Miller. Loeb Edition. (Harvard University Press, II, xxii, 79, Cambridge, 1913).</bok>

<eref>2. L'[Académie des sciences de l'Institut de France, Du bon usage de la bibliométrie pour l'évaluation individuelle des chercheurs, available at http://www.sauvonsluniversite.com/spip.php?article4391 (2011).</eref>

<jrn>3. E. Garfield, Citation indexes for science: A new dimension in documentation through association of ideas. *Science* **122**, 108–111 (1955).doi:10.1126/science.122.3159.108 Medline</jrn>

<eref>4. L'Académie des sciences de l'Institut de France, Évaluation des chercheurs et des enseignants-chercheurs en sciences exactes et expérimentales: Les propositions de l'Académie des sciences, available at: http://www.academie-sciences.fr/actualites/textes/recherche_08_07_09.pdf (2009).</eref>

<eref>5. The Thomson Reuters Impact Factor, available at http://thomsonreuters.com/products_services/science/free/essays/impact_factor/.</eref>

<jrn>6. B. M. Althouse, J. D. West, C. T. Bergstrom, T. Bergstrom, Differences in impact factor across fields and over time. *J. Am. Soc. Inf. Sci. Technol.* **60**, 27–34 (2009). doi:10.1002/asi.20936</jrn>

<jrn>7. P. O. Seglen, Why the impact factor of journals should not be used for evaluating research. *BMJ* **314**, 498–502 (1997). Medline</jrn>

<jrn>8. D. W. Aksnes, A macro-study of self-citation. *Scientometrics* **56**, 235–246 (2003). doi:10.1023/A:1021919228368</jrn>

<jrn>9. J. E. Hirsch, An index to quantify an individual's scientific research output. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 16569–16572 (2005).doi:10.1073/pnas.0507655102 Medline</jrn>

<jrn>10. A. Sidiropoulos, D. Katsaros, Y. Manolopoulos, Generalized Hirsch *h*-index for disclosing latent facts in citation networks. *Scientometrics* **72**, 253–280 (2007). doi:10.1007/s11192-007-1722-z</jrn>

<jrn>11. M. E. Falagas, E. I. Pitsouni, G. A. Malietzis, G. Pappas, Comparison of PubMed, Scopus, Web of Science, and Google Scholar: Strengths and weaknesses. *FASEB J.* **22**, 338–342 (2008).doi:10.1096/fj.07-9492LSF Medline</jrn>

<jrn>12. A. V. Kulkarni, B. Aziz, I. Shams, J. W. Busse, Comparisons of citations in Web of Science, Scopus, and Google Scholar for articles published in general medical journals. *JAMA* **302**, 1092–1096 (2009).doi:10.1001/jama.2009.1307 Medline</jrn>

<edb>13. A. W. Harzing, in *The Publish or Perish Book* (Tarma Software Research Pty Ltd, Melbourne, Australia, 2010).</edb>

<jrn>14. L. Egghe, Theory and practice of the *g*-index. *Scientometrics* **69**, 131–152 (2006). doi:10.1007/s11192-006-0144-7</jrn>

<jrn>15. D. Rosenstreich, B. Wooliscroft, Measuring the impact of accounting journals using Google Scholar and the *g*-index. *Br. Account. Rev.* **41**, 227–239 (2009) (BAR). doi:10.1016/j.bar.2009.10.002</jrn>

<jrn>16. N. Haeffner-Cavaillon, C. Graillot-Gak, The use of bibliometric indicators to help peer-review assessment. *Arch. Immunol. Ther. Exp. (Warsz.)* **57**, 33–38 (2009).doi:10.1007/s00005-009-0004-2 Medline</jrn>

<jrn>17. A. Abbott, D. Cyranoski, N. Jones, B. Maher, Q. Schiermeier, R. Van Noorden, Metrics: Do metrics matter? *Nature* **465**, 860–862 (2010).doi:10.1038/465860a Medline</jrn>