Metrics of Science

January 01, 2008

Assessments of science are important for many different reasons. For individuals early in their careers, metrics of scientific work can provide valuable feedback about where they stand and the progress they have made. For faculty seeking to hire another member of their department, such metrics can simplify the task of wading through hundreds of applications to identify a subset of applicants to interview. For departmental chairs, these metrics may influence annual raises and the allocation of scarce departmental resources. For university administrators, these metrics help identify faculty who warrant promotion and tenure. For scientific societies, these metrics influence the selection of award recipients across the course of careers. For funding agencies, both public and private, assessments of science help identify areas of progress and vitality that may warrant additional resources. For legislative bodies and boards of directors, measures of science provide a means of documenting performance, ensuring accountability, and evaluating the return on their research investment. Measures of science also can be used for a variety of other purposes, such as identifying the structure of science, the impact of academic journals, influential fields of research that warrant funding or support, and factors that may contribute to the likelihood of discoveries.

The simplest metric is scholarly productivity or number of publications, and this metric has been found to be a good predictor of career trajectories (Simonton, 1997). There are numerous variants on this metric, such as the number of peer-reviewed articles, the number of first-authored publications, the number of articles published in a specific premier journal in a field, the number of articles without more senior collaborators, and the number of publications across years (publication trajectories).

James Byrnes (2007) recently examined the publication trends of psychology faculty during their pretenure years. He began by noting several publishing trends in the discipline, such as faculty in highly-ranked departments publishing at nearly twice the rate as faculty at lower-ranked departments. Byrnes (2007) then limited his analysis to the publishing trends of faculty who had earned their doctorates between 1973 and 1999 and who currently hold tenured appointments at highly ranked psychology departments. He found that, on average, the faculty published 11.03 peer-reviewed articles in their first seven post-doctoral years. Analyses of the trajectory of publishing revealed the rate of publication increased each year following completion of the doctorate, with the largest increases in the rate of publication of peer-reviewed journal articles observed in the first four years rather than in the two years immediately before tenure. Although Byrnes (2007) did not report the publication rate of those denied tenure in these departments, there was no evidence that, on average, faculty who were tenured engaged in a “tenure push.”

Publication counts prior to tenure do not tell the whole story, of course. In an investigation of gender differences in scientific productivity, Long (1992) found that women published fewer articles than men during the first decade of their career, but that this difference was reversed later in their careers.

It comes as news to no one that if a candidate for tenure has no or very few scientific publications, the

assessment of the candidate's science is likely to be bleak. But if few publications provide little evidence for tenure, it does not follow that more publications are necessarily better. The quality, innovativeness, programmatic nature, and cumulative impact of the research are among the features that might be important to consider as well. Traditionally, additional metrics were garnered from the prestige of the journal in which the work was published and the order of authorship. But forgettable or flawed scientific work appears in prestigious outlets, the last author in some fields (e.g., the neurosciences) is as prestigious or important as the first, and scientific knowledge increasingly is being advanced by collaborative scientific teams. Therefore, a faculty's reading of a body of scientific work, scientific presentations and discussions with an author or authors, and the views expressed by external experts have served as additional measures of a body of scientific work. The reasoning is that the opinions of experts in a field matter, and the judgment of expert reviewers is the basis of our cherished, if imperfect, peer review system. But the subjective nature of these appraisals, concerns about the role of the body politic, occasionally low inter-judge agreement, their insensitivity to variations in activity across time, the possibility of external reviewers no longer being willing to provide frank and honest evaluations given that their letters of evaluation may not be kept confidential, and difficulties in comparing such assessments across different fields may leave one searching for more objective alternatives. The need for such alternatives is especially evident when measures of science are used in macro-analyses, such as those used to determine influential fields of research that warrant support or to provide a means of evaluating the return on research investments.

More than three decades ago, a senior colleague suggested that people should be allowed to publish any paper or papers they wish. When the look of horror on our faces became apparent, an outcome that was measurable in milliseconds, the individual explained that there would be so much published that counting would become meaningless and that readers would be attracted to particular authors or bodies of work based on innovativeness, import, and quality. This individual presaged the notion underlying contemporary citation impact analyses.

By indexing reference lists of articles published in journals, the Institute for Scientific Information (ISI) has made it simple to determine the number of times particular articles, bodies of work by an author, or articles appearing in a particular journal were cited in published journal articles. Although errors can occur, indexing citations is generally objective and reliable, and it can be calculated (excluding self-citations) to provide a metric of the extent to which a work or body of work has influenced others. Although the citation impact metric is commonly described as a measure of the scientific influence an article, author, or journal, it is commonly treated as a measure of scientific contribution or merit.

Limitations of the citation count have been detailed previously (e.g., Hébert, 2004; Roediger, 2006): articles appearing in early issues of new journals may not be indexed; books, chapters, and published conference proceedings are seldom or irregularly indexed; some authors' work becomes so well-accepted that they no longer are cited for the discovery; errors in references ranging from the misspelling of author names to the misspecification of the page numbers or author initials contribute to underestimates of citation impact; authorship order is ignored; tutorials of methodological or statistical techniques may garner large citation counts compared to the groundbreaking work that led to the development of the technique; articles in faddish or insulated areas may have a sizable short-term influence, but little if any long-term influence; insulated groups of investigators can collude to cite each other's work as a means of inflating citation counts; fields differ in terms of conventions concerning references to existing literature, half-lives for published work, and lags between submission of

manuscripts; and publication of articles varies across journals, fields, and disciplines.

Limitations in the ISI dataset (Web of Knowledge), such as the typical absence of books and chapters, may be filled by other citation databases, including Scopus and Google Scholar. Scopus provides better coverage of conferences than Web of Knowledge, but the coverage of publications prior to 1992 is poor. Google Scholar has good coverage of conferences, books, book chapters, and most journals but the coverage of journals prior to 1990 is somewhat limited and work that is not available through archival publication channels (gray literature) is included. This broader coverage usually leads to higher citation counts using Google Scholar rather than Web of Knowledge for searches in the social sciences, humanities, and engineering. At this juncture, Google Scholar's coverage of some fields of science is less complete than ISI. In addition, some publishers do not permit open access to their journals for one year, which delays their availability through public search engines such as Google Scholar. Google Scholar and Scopus, therefore, are best regarded currently as a complement to rather than a replacement for Web of Knowledge. (APS has been in discussions with Google to ensure that the journals most relevant to psychological scientists are covered.)

Given a good source of publication and citation data, it is a simple matter to calculate the total number of publications, total number of citations, average number of citations per paper, average number of citations per author, average number of papers per author, average number of publications per year, and average number of citations per year. In addition, other metrics of science have been proposed to address some of the problems with total citation counts. One such index, which has been discussed previously in the *Observer*, is Hirsch's h-index (Hirsch, 2005; Roediger, 2006). The h-index is defined as the number of papers (*h*) authored by an individual (or group of individuals, department, university, country, etc) with citations equal to or greater than *h*.

The h-index provides a single number that balances the number of publications and the number of citations per publication. John Ridley Stroop, for instance, published his dissertation as an article entitled "Studies of interference in serial verbal reactions" in the *Journal of Experimental Psychology* in 1935. After attaining his PhD in 1933 from George Peabody College and serving briefly as a special instructor in psychology and education at Tennessee Polytechnic Institute, he returned to David Lipscomb University where he served in various capacities (including as registrar and professor of biblical studies). According to a search of the ISI database, Stroop published a total of three papers in the field during his career. The article based on his dissertation has been cited 3,810 times, and the other two papers have had less than 1 percent of this citation impact. Total number of citations is relevant to an evaluation of this person's scientific merit, but it misses the point that Stroop's scientific contributions to psychology were limited primarily to his efforts prior to the completion of his PhD. John Stroop's h-index is 3, which places his total citation count into perspective. As this example illustrates, the citation count *and* the h-index provide important information not contained in either alone, so using both metrics can be preferable to using either alone.

The h-index also has limitations. It tends to increase with years as a scientist; gratuitous authorship can contribute to inflated scores; the h-index does not take into consideration the number or the role of the authors; different citation databases provide different h-indexes as a result of differences in coverage; the h-index is bounded by total number of publications; the h-index does not consider the context of the citations (e.g., negative findings or retracted work); and individuals with the same h-index may nevertheless differ dramatically in total citations or in number of publications. The g-index, proposed by

Leo Egghe (2006), attempts to address the last concern by giving more weight to highly cited articles. Given a set of articles ordered in terms of decreasing citation counts, the g-index is the largest number such that the top *g* articles together total at least g^2 citations. The g-index will be equal to or greater than the h-index. In the case of Stroop, the g-index equals the h-index because, based on ISI citation data, his total number of published articles is 3.

If what one seeks to gauge is recent impact, the contemporary h-index is useful (Sidiropoulos, Katsaros, & Manolopoulos, 2006). This index allows the user to designate what temporal parameters to use, but the notion is to weight the citation impact of each article based on years since it was published. A parameter of 5 and delta of 1, for instance, means that citations to articles published in the current year are multiplied by 5, total citations to an article published five years ago are multiplied by 1, articles published 10 years ago are multiplied by .5, and so forth. The contemporary h-index is superior to the h-index in identifying who has remained productive and influential or, if calculated for topics rather than individuals, what topics are hot.

A tool provided by Google, *Publish or Perish*, provides these indices and more. The program is free and can be downloaded from http://www.harzing.com/resources.htm#/pop.htm. The output provides total number of papers, total number of citations, average number of citations per paper, average number of citations per author, average number of papers per author, average number of citations per year, analysis of the number of authors per paper, Hirsch's h-index and related parameters, Egghe's g-index, the contemporary h-index, h-indices that measure the per-author impact of articles (individual h-indices), and variations on age-weighted citation rates that measure the total number of citations to a body of work adjusted for the age of each individual paper.

Which indices are preferred depends on the question that is asked. No single index provides an optimal metric of science, whether scaled at the level of the individual scientist, topic, field, journal, or discipline. For instance, citation data may not be as helpful at research universities when evaluating candidates for the position of beginning assistant professor as when evaluating candidates for promotion to full professor. Moreover, as in the case of operationalizing any theoretical variable, using multiple metrics of science has advantages. When multiple metrics of science converge on the same result, confidence in the result is increased. When these metrics provide different outcomes, as in the case of John Ridley Stroop, then discrepancies across the indices can be used to draw more informed interpretations.

The availability of quantitative data and analysis tools like *Publish or Perish* makes metric-based decision making simpler than ever. As objective and replicable as the results might appear (and as valuable as they might appear to be), a caveat is in order. Metric-based decision making can have the unintended effect of promoting scientific work that yields higher values on the selected metric rather than more meaningful or innovative scientific work. It is worth emphasizing that none of these measures is a perfect index of scientific merit (and none of the citation databases are perfect sources of data). For this reason, substantive consultations, deliberations, and evaluations (e.g., by search committees, chairs, deans, peer review) continue to have an important place in the measurement of science. Pierre Wiltzius, Director of the Beckman Institute for Advanced Science and Technology at the University of Illinois at Urbana-Champaign, told me recently that in addition to quantitative metrics and consultations with an external advisory committee, he makes it a point to personally familiarize himself with the work of the faculty, and he uses a social network analysis to identify which scientists positively impact the work of

each scientist. The network analysis helps to identify the engines of their scientific enterprise and important bridge scientists without whom the total would be less than the sum of the parts. ?

References

Byrnes, J.P. (2007). Publishing trends of psychology faculty during their pretenure years. *Psychological Science, 18*, 283-286.

Feller, I., & Stern, P.C. (2007). *A strategy for assessing science*. Washington, DC: National Academies Press.

Egghe, L. (2006) Theory and practice of the g-index, *Scientometrics, 69*, 131-152.

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

Long, J.S. (1992). Measures of sex differences in scientific productivity. *Social Forces, 71*, 159-178.

Roediger, III, H.L. (2006). The h index in science: A new measure of scholarly contribution. *Observer, 19*(4).

Hébert, R. (2004). Highly cited, highly controversial. *Observer, 17*(3).

Sidiropoulos, A., Katsaros, D., & Manolopoulos, Y. (2007). Generalized h-index for disclosing latent facts in citation networks. *Scientometrics, 72*, 253-280.

Simonton, D.K. (1997). Creative productivity: A predictive and explanatory model of career trajectories and landmarks. *Psychological Review, 104*, 66-89.

Stroop, J.R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology, 18*, 643-662.