Journal of Informetrics 11 (2017) 595–597

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

Book review

## In Praise of Bibliometrics

**Bibliometrics and Research Evaluation: Uses and Abuses, Yves Gingras, MIT Press, Cambridge, MA (2016). 119 + xii pages, Hardcover US$ 26; Paperback US$ 18**

In a concise book entitled *Bibliometrics and Research Evaluation: Uses and Abuses*, Yves Gingras argues for the correct use of bibliometric indicators. The book is comprehensive, essayistic, and partisan in defense of the use and potential of citation analysis.

In Chapter 1, Gingras recalls the history of bibliometrics first as a management tool for librarians in the 1920s until the creation of the Science Citation Index (SCI) in the 1950s. The SCI then became a tool in the hands of historians and philosophers of science in the 1960s and 70s and science policy in the 1970s and finally in research evaluation in the following decades. In a spirit comparable to the one when bibliometrics emerged as a specialty—in response to Garfield's introduction of the SCI (Cole & Cole, 1973; Elkana, Lederberg, Merton, Thackray, & Zuckerman, 1978; Price, 1965, 1976)—Gingras considers citation indexes as rich resources which enable historians and philosophers of science to refine their analyses and substantiate their claims empirically (Chapter 2).

In Chapter 3 entitled "The Proliferation of Research Evaluation," Gingras turns to the uses and abuses of bibliometric indicators in research evaluation. In a fourth chapter entitled "The Evaluation of Research Evaluation," the author criticizes the "abuses" of rankings by university managers, and notably at Business Schools, as "intellectual corruption" (p. 83). The bibliometric enterprise, however, is to be rescued by denouncing these abuses as based on misunderstandings and sometimes plain error.

It is often a pleasure to see how Gingras analyzes and criticizes. Because of the relatively recent use of citation-based indicators for evaluation and ranking (since the mid-1990s), a host of misunderstandings and errors have to be explained and corrected. Where necessary, the author adds his own bibliometric analyses in order to show how abuses and error can and should be corrected. The top-rank of the King Abdulaziz University in Saudi Arabia is mentioned (pp. 84f.), but not fully explained, in my opinion. How does it work? How much are top-authors paid for adding this affiliation to their name? One could ask a few of these authors about the details.

Commenting on the use of the third decimal for Journal Impact Factors (JIFs), Gingras formulates (p. 49): "Misuse of rankings and indicators with a patina of precision basically betrays ignorance of the properties of the indicators used. I don't need to comment in more detail here on the many perverse effects generated by the use of flawed indicators." Gingras' solutions are technical and rational. For example, he argues that the incompatibility of using JIFs across different fields of science would disappear if one used 10-year JIFs. Differences in citation turn-over between fields with and without competition at research fronts (Price, 1970) would disappear using a sufficiently large time window. However, these arguments ignore that most users are less knowledgeable about the choices of parameters in the model than professional scientometricians and are often not interested in the technical details. As long as the indicators can legitimately be used for making decisions, most users continue to work with three decimals and short-term JIFs despite the warnings of professionals. Gingras considers these bureaucratic uses of "black-boxed" indicators as illegitimate and incorrect; he engages in discussions. But the logic in his reasoning remains rationalistic and technical. The logic in other domains is different: management cannot wait ten years before making a decision. R&D managers and S&T policy makers can work with erroneous statistics as long as these are not delegitimated.

Gingras argues in defense of using citations as robust at sufficiently high levels of aggregation. Major trends are reliably indicated, but yearly rankings, for example, should be mistrusted since based on a single measurement point. In his opinion, one can always focus on the exceptions at the margins and find problems; but these problems can be clarified as based on misunderstanding and error. But is answering the criticism so easy? Is not the JIF which Gingras defends (!) itself an average based on a skewed distribution and therefore already a source of error? Isn't the problem that in an evaluation one always needs to define classes—groups and categories—an unavoidable source of bureaucratically generated error? The delineations

http://dx.doi.org/10.1016/j.joi.2017.03.002
1751-1577/

are fuzzy in terms of the data, and one risks reifying differences where they are not sharp. This seems unavoidable to me: one needs reference sets in order to be able to compare across different sets. The reference sets may be insufficiently homogenous (e.g., the Information and Library Science category in the Web of Science; see Leydesdorff & Bornmann, 2016), and the sets to be evaluated can be "interdisciplinary" and may therefore not fit in the boxes (e.g., Rafols, Leydesdorff, O'Hare, Nightingale, & Stirling, 2012). In other words, "error" is unavoidable. Is Gingras fighting an uphill battle? On the wrong hill?

Let me take my own position as an example (because I don't wish to harm anyone by an evaluation at the individual level). As an information scientist/scientometrician in a department of communication studies, my papers don't fit into the grouping that is needed for institutional evaluation. The problem is not so easily solvable by abstracting intellectually from it: the error is persistent. If one were to solve it by reorganizing the university departments for this type of reasons, one would find, for example, that not all my articles are classifiable as information science; some of them are more sociological. In other words, the aggregation along institutional or intellectual lines generates different types of error.

An evaluation requires a model; and a model can be expected to have effects on the ranking. The effects of the model are unavoidable, and evaluees may therefore legitimately complain and propose other models. To rely on trends at the aggregated level assumes that error "averages out" where numbers are sufficiently large. Are universities, for example, sufficiently large for this purpose? The robust trends at high levels of aggregation may be useful for informing S&T policies at the national level, but institutional evaluation requires detail and precision in categorizing at the bottom. One usually evaluates a cell in a matrix of institutional and disciplinary structures; for example, the faculties of chemistry in a country. But one cannot compare the Faculty of Chemistry in University A with its counterpart in University B without taking into account how chemistry is sub-divided among specialties within chemistry (e.g., biochemistry versus physical chemistry). Furthermore, relevant chemistry may also be done in the biology faculty or the faculty of medicine.

Gingras does not address these problems in evaluation practices because, in his opinion, they can be overcome by rational discourse, correcting mistakes and abuses. However, I am not so sure about this: the evaluation models contain parameter choices which affect the results. Recently, for example, we found that the decline of Carnegie Mellon University in the Leiden Rankings between 2013 and 2016 was due for more than 72 percent to the model (Leydesdorff, Wouters, & Bornmann, 2016: pp. 2144f.). It was possible to specify this "error" because the analysis underlying the Leiden Rankings is transparent and error-free (given the state of the art).

In my opinion, the practices of bibliometrics have changed the status of the field since the 1970s. Bibliometrics has become the subject of a political economy that it co-constructs. When the problems are not epistemic but political or social, they are serious problems nevertheless. At some places, Gingras admits this. For example, the *h*-index should not be used, since the consistency problems with it cannot be solved intellectually (Waltman & Van Eck, 2012). The Shanghai Rankings are invalid because based on a summation of six different measures: "Clearly, the final index is not adequate since it is a composite of several heterogeneous measures: the number of publications in *Science* and *Nature* is not commensurate with the number of Nobel Prizes" (p. 77). However, professional bibliometricians have to make practical assumptions.

For example, Gingras and Larivière (2011) themselves advocated using what was later renamed the Mean Normalized Citation Score (MNCS). Are there no problems with using this indicator? Does not taking the mean of a skewed distribution generate error (Leydesdorff, Bornmann, Mutz, & Opthof, 2011)? Some of us have now changed to using percentiles as a non-parametric alternative. But there are no theoretical reasons for choosing the one or the other set of percentile classes, whereas the effects of using the top-10% or the top-25% for an evaluation can be large (Perianes-Rodríguez & Ruiz-Castillo, 2016).

Let me emphasize that I don't have answers to the questions which I raise; but I don't think Gingras has them either. Our open questions shape a research agenda. Perhaps it makes the field vulnerable to critique more than arguing against criticism as ill-informed and mistaken. Is one able to engage in discussions with "users and abusers" with other standards, practices, and objectives? This book addresses an external audience of policy-makers, R&D managers, and practicing scientists in defense of bibliometrics; and from this perspective it provides a commendable argument.

## References


Cole, J. R., & Cole, S. (1973). *Social stratification in science*. Chicago/London: University of Chicago Press.

Elkana, Y., Lederberg, J., Merton, R. K., Thackray, A., & Zuckerman, H. (1978). *Toward a metric of science: The advent of science indicators*. New York, etc: Wiley.

Gingras, Y., & Larivière, V. (2011). There are neither king nor crown in scientometrics: Comments on a supposed alternative method of normalization. *Journal of Informetrics*, *5*(1), 226–227.

Leydesdorff, L., & Bornmann, L. (2016). The operationalization of fields as WoS subject categories (WCs) in evaluative bibliometrics: The cases of library and information science and science & technology studies. *Journal of the Association for Information Science and Technology*, *67*(3), 707–714.

Leydesdorff, L., Bornmann, L., Mutz, R., & Opthof, T. (2011). Turning the tables in citation analysis one more time: Principles for comparing sets of documents. *Journal of the American Society for Information Science and Technology*, *62*(7), 1370–1381.

Leydesdorff, L., Wouters, P., & Bornmann, L. (2016). Professional and citizen bibliometrics: Complementarities and ambivalences in the development and use of indicators—A state-of-the-art report. *Scientometrics*, *109*(3), 2129–2150.

Perianes-Rodríguez, A., & Ruiz-Castillo, J. (2016). *The impact of classification systems in the evaluation of the research performance of the Leiden Ranking Universities*. http://e-archivo.uc3m.es/handle/10016/22147

Price, D. J. de Solla. (1965). Networks of scientific papers. *Science*, *149*(3683), 510–515.

Price, D. J. de Solla. (1970). Citation measures of hard science, soft science, technology, and nonscience. In C. E. Nelson, & D. K. Pollock (Eds.), *Communication among scientists and engineers* (pp. 3–22). Lexington, MA: Heath.

Price, D. J. de Solla. (1976). A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science*, *27*(5), 292–306.

Rafols, I., Leydesdorff, L., O'Hare, A., Nightingale, P., & Stirling, A. (2012). How journal rankings can suppress interdisciplinary research: A comparison between innovation studies and business & management. *Research Policy*, *41*(7), 1262–1282.
Waltman, L., & Van Eck, N. J. (2012). The inconsistency of the h-index. *Journal of the American Society for Information Science and Technology*, *63*(2), 406–415.

Loet Leydesdorff
*University of Amsterdam, Amsterdam School of Communication Research (ASCoR), P.O. Box 15793, 1001 NG Amsterdam, The Netherlands*
*E-mail address:* loet@leydesdorff.net
Available online 18 May 2017