



STI 2018 Leiden

*23rd International Conference on Science and Technology Indicators
"Science, Technology and Innovation Indicators in Transition"*

STI 2018 Conference Proceedings

Proceedings of the 23rd International Conference on Science and Technology Indicators

All papers published in this conference proceedings have been peer reviewed through a peer review process administered by the proceedings Editors. Reviews were conducted by expert referees to the professional and scientific standards expected of a conference proceedings.

Chair of the Conference

Paul Wouters

Scientific Editors

Rodrigo Costas
Thomas Franssen
Alfredo Yegros-Yegros

Layout

Andrea Reyes Elizondo
Suze van der Luijt-Jansen

The articles of this collection can be accessed at <https://hdl.handle.net/1887/64521>

ISBN: 978-90-9031204-0

© of the text: the authors

© 2018 Centre for Science and Technology Studies (CWTS), Leiden University, The Netherlands



This ARTICLE is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Responsible metrics: One size doesn't fit all¹

Ludo Waltman

waltmanlr@cwts.leidenuniv.nl

Centre for Science and Technology Studies (CWTS), Leiden University, P.O. Box 905, 2300 AX Leiden (the Netherlands)

Introduction

Responsible use of scientometrics in research evaluations is heavily debated. In recent years, a number of high-profile statements on 'responsible metrics' were published, most notably the San Francisco Declaration on Research Assessment (DORA, 2013), the Leiden Manifesto (Hicks, Wouters, Waltman, De Rijcke, & Rafols, 2015), and the Metric Tide report (Wilsdon et al., 2015). Each of these statements presents a number of principles for responsible use of scientometrics in research evaluations. These principles have been widely discussed, and they have inspired several organizations to develop guidelines for the use of scientometrics in the evaluations they perform. At the same time, the principles presented in the above-mentioned statements are quite general, and it is therefore not always clear how they can be applied in a specific evaluative setting.

In this opinion paper, my aim is to draw attention to the importance of distinguishing between different evaluative settings. Principles for responsible use of scientometrics are dependent on the nature of the evaluative setting. What is responsible in one setting may not be responsible at all in a different setting. Recognizing that there is no one size that fits all leads to insights that may help to enhance responsible use of scientometrics. To illustrate the importance of distinguishing between different evaluative settings, I will focus on the distinction between micro-level and macro-level research evaluation.

Micro-level vs. macro-level research evaluation

Research evaluation may refer to many different things. Some of us may think primarily of evaluations of individual researchers in job interviews, tenure procedures, and grant applications. Others may think of evaluations of research groups and university departments in local or national evaluation exercises. Yet others may have in mind evaluations of universities, for instance by governments that need to allocate funding, or even evaluations of entire countries from an international comparative perspective. In this paper, I distinguish between micro-level and macro-level research evaluation. Micro-level evaluation refers to the in-depth evaluation of individual researchers and research groups. Macro-level evaluation is for instance about the evaluation of entire research institutions and countries.

The idea of combining scientometric indicators with expert assessment plays a key role in the responsible metrics debate. The Leiden Manifesto for instance states that "quantitative evaluation should support qualitative, expert assessment", and this is also one of the guiding

¹ This paper is an extended version of a blog post available at www.cwts.nl/blog?article=n-r2s294.

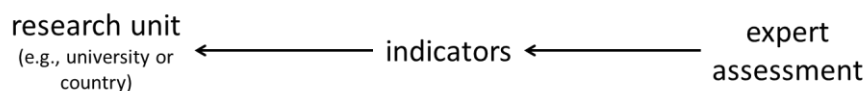
principles put forward in the Metric Tide report. However, as I will argue below, the relation between scientometric indicators and expert assessment differs fundamentally between the micro and the macro level (see also Figure 1), leading to different requirements for responsible use of scientometrics at these two levels.

Figure 1: Illustration of the distinction between micro-level and macro-level research evaluation. At the micro level, peer review by experts is the recommended approach to research evaluation. Experts are able to directly evaluate a micro-level research unit (e.g., a researcher or a research group). They may be supported by indicators that summarize potentially useful information about the performance of the research unit. At the macro level, experts are unable to perform an in-depth evaluation of an entire research unit (e.g., a university or a country). Instead, they need to rely on indicators that provide aggregate information about the performance of the research unit.

Micro-level research evaluation



Macro-level research evaluation



Micro-level research evaluation

At the micro level, scientometricians generally recommend to rely on expert assessment for an in-depth evaluation of for instance an individual researcher. The crucial role of expert assessment, or peer review, in micro-level research evaluation is emphasized in the Leiden Manifesto, which recommends to “base assessment of individual researchers on a qualitative judgement of their portfolio”. Although scientometric indicators may provide valuable information, they are unable to provide a comprehensive perspective on the performance of a researcher. In a micro-level research evaluation, these indicators therefore should be used only to support expert assessment. Indicators may for instance be employed to summarize potentially useful information about the performance of a researcher and to draw attention to remarkable accomplishments of a researcher. The use of scientometric indicators to support expert assessment is sometimes referred to as informed peer review.

Macro-level research evaluation

No evaluator has a comprehensive overview of all research activities taking place at an institution or in a country, let alone the expertise needed to perform an in-depth evaluation of all these activities. At the macro level, the value of expert assessment therefore is not in the detailed evaluation of individual research activities, but in combining – in a balanced and well-informed manner – many different pieces of information that together provide the big

picture of macro-level research performance. Some of these pieces of information may be provided by qualitative indicators, resulting for instance from peer review carried out at lower levels of aggregation, while other information may be obtained from quantitative indicators, for instance indicators based on bibliometric data.

The macro level differs fundamentally from the micro level. At the macro level, experts view the world through indicators. Viewing the world directly, without using indicators, is almost impossible at this level. This means that macro-level research evaluation cannot be done without indicators. Hence, indicators do not just support research evaluation at the macro level. They enable macro-level research evaluation.

Having discussed the distinction between micro-level and macro-level research evaluation, I will now turn to the implications of this distinction for the debate about responsible use of scientometrics.

Sophistication vs. simplicity

Scientometricians put a lot of effort into designing increasingly sophisticated indicators (e.g., Waltman, 2016). However, at the same time, they argue that indicators should be simple and transparent. This tension is for instance visible in the Leiden Manifesto, which states that “normalized indicators are required” to “account for variation by field in publication and citation practices”, while it also recommends that analytical processes should be kept “open, transparent and simple”. Almost inevitably, a normalized indicator is less simple and less transparent than a non-normalized indicator, creating a tension between different recommendations made by the Leiden Manifesto. The distinction between micro-level and macro-level research evaluation provides guidance on how to deal with this tension.

At the macro level, experts view the world through indicators. Given the strong dependence of experts on indicators, it is essential that each indicator has an unambiguous conceptual foundation and a high degree of validity. Scientometricians often emphasize the importance of these properties of indicators (e.g., Gingras, 2016, p. 71–77; Sugimoto & Larivière, 2018, p. 14–16). In order to guarantee a high level of validity, indicators typically need to be relatively sophisticated at the macro level. Simple indicators often do not have the required level of validity, and these indicators may therefore provide experts with a distorted world view. For instance, in a research evaluation that covers multiple fields, a simple citation-based indicator of scientific impact that does not include a field normalization has a low validity, because it does not correct for differences between fields in citation practices. Such an indicator may give the incorrect impression that research units active in fields with a high citation density (e.g., in the life sciences) are more impactful than their counterparts active in fields with a low citation density (e.g., in the social sciences). In order not to provide experts with such a biased world view, a more sophisticated impact indicator should be used that includes a normalization for field differences.

However, a high level of sophistication comes at a cost. Highly sophisticated indicators tend to become black boxes. It is difficult to understand the internal workings of these indicators. This is problematic especially in micro-level research evaluation. At the micro level, indicators are intended to support expert assessment in a process of informed peer review. As discussed above, indicators may be used to summarize potentially useful information and to draw attention to remarkable accomplishments. Indicators do not so much provide information themselves, but they point experts to information that may be of relevance and that may need to be examined in more detail. Complex black box indicators are difficult to

use in such a process. The problem of a black box indicator is that it is almost impossible to go back from the indicator to the underlying information. For instance, when an indicator has a surprisingly high value, experts need to understand how this high value can be explained. However, this is difficult when working with a black box indicator. Is the high value of the indicator due to an idiosyncratic outlier, is it an artifact of the algorithm used to calculate the indicator, or does it result from a genuinely outstanding research performance? In the case of a black box indicator, experts can only guess. Keeping indicators simple ensures that experts can truly reflect on what indicators tell them and can take this into account in their expert assessment. This is an important argument for using simple indicators at the micro level.

Professional vs. citizen scientometrics

The distinction between micro-level and macro-level research evaluation also provides insight into the tension between professional scientometrics and citizen scientometrics (Hammarfelt & Rushforth, 2017; Leydesdorff, Wouters, & Bornmann, 2016). Citizen scientometricians perform scientometric analyses without paying much attention to the body of knowledge accumulated in the professional scientometric community. They develop their own ad hoc scientometric solutions, for instance by making use of readily available tools such as the journal impact factor, the h-index, and Google Scholar. Professional scientometricians tend to be critical of citizen scientometrics, arguing that citizen scientometricians perform low-quality scientometric analyses. However, the distinction between micro-level and macro-level research evaluation suggests that there is a need to develop a more nuanced perspective on citizen scientometrics.

In macro-level research evaluation, concerns about citizen scientometrics seem warranted. Solutions developed by citizen scientometricians most likely do not have the required level of sophistication. Scientometric analyses at the macro level should therefore be based on the ideas and standards developed in the professional scientometric community. However, the situation is different in micro-level research evaluation. At the micro level, scientometric indicators are intended to support expert assessment in a process of informed peer review. Experts are expected to have a detailed knowledge of the characteristics of a specific evaluative setting. To make sure that indicators are helpful in supporting expert assessment, this knowledge of experts needs to be taken into account in the choice and the design of indicators. In order to accomplish this, experts carrying out a micro-level research evaluation should be fully engaged in selecting the most relevant indicators and in fine-tuning these indicators to the peculiarities of the evaluative setting. This not only has the advantage that indicators will be truly meaningful and relevant to experts, but it also ensures that experts are able to reflect in detail on what indicators tell them. Engaging experts in the choice and the design of indicators requires experts to take on a role as citizen scientometrician. Hence, to realize the idea of informed peer review in micro-level research evaluation, experts need to become citizen scientometricians. Unlike citizen scientometricians, professional scientometricians lack an in-depth understanding of a specific evaluative setting. Their role at the micro level should therefore focus on providing general guidance on the use of indicators. Professional scientometricians may for instance provide advice on scientometric data sources that could be used, they may suggest indicators that could be of relevance, and they may offer help in integrating indicators in a process of informed peer review.

Indicators vs. statistics

Another implication of the distinction between micro-level and macro-level research evaluation relates to the nature of scientometric indicators at the two levels. As discussed above, at the macro level, experts view the world through indicators, and it is therefore

essential for indicators to have an unambiguous conceptual foundation and a high degree of validity. In contrast, at the micro level, simplicity is a more important property of indicators. My suggestion is to adopt a terminology that reflects this difference between the two levels. While ‘indicators’ should remain the preferred term at the macro level (‘metrics’ is also sometimes used, but ‘indicators’ is often considered more appropriate), at the micro level I propose to use ‘statistics’ instead of ‘indicators’. Indicators aim to provide a proxy of an underlying concept. Statistics are less ambitious and are not necessarily expected to represent a well-defined concept. They just offer a quantitative summary of a certain body of information. The use of statistics that do not have a clear interpretation in terms of a well-defined concept (e.g., many altmetric statistics) might seem dangerous, but it is important to keep in mind that these statistics are intended to be embedded in a process of informed peer review. By refraining from imposing rigid interpretations on statistics, experts are given full freedom to interpret statistics in the light of their expert understanding of a specific evaluative setting.

Conclusion

How scientometrics can be used responsibly depends on the nature of the evaluative setting. The distinction between micro-level and macro-level research evaluation serves as an illustration of this point. There is a need for sophisticated indicators at the macro level, while simple indicators are preferable at the micro level (as is also emphasized by the idea of contextualized scientometrics; see Waltman & Van Eck, 2016). Also, the involvement of professional scientometricians is essential at the macro level, while citizen scientometricians can play an important role at the micro level.

Of course, things are more complex than a simple micro-macro distinction may suggest. There is a meso level (e.g., evaluation of departments or institutes within a university) at which one needs to strike a balance between the recommendations made for the micro and macro levels. Also, at each level (i.e., micro, meso, or macro), there is a considerable heterogeneity in different types of evaluations. Furthermore, there are interdependencies between the various levels, with evaluations at one level influencing evaluations at other levels.

More refined taxonomies of different types of evaluations may lead to further differentiation in recommendations for responsible use of scientometrics. To make progress in the responsible metrics debate, we need to recognize that there is no one size that fits all.

Acknowledgment

I would like to thank Rodrigo Costas, Sarah de Rijcke, Ismael Rafols, and Paul Wouters for valuable feedback on earlier versions of this paper.

References

DORA (2013). *San Francisco declaration on research assessment*. Retrieved April 9, 2018, from <https://sfdora.org>

Gingras, Y. (2016). *Bibliometrics and research evaluation: Uses and abuses*. MIT Press.

Hammarfelt, B., & Rushforth, A.D. (2017). Indicators as judgment devices: An empirical study of citizen bibliometrics in research evaluation. *Research Evaluation*, 26(3), 169–180.

Hicks, D., Wouters, P., Waltman, L., De Rijcke, S., & Rafols, I. (2015). The Leiden Manifesto for research metrics. *Nature*, 520, 429–431.

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.

Sugimoto, C.R. & Larivière, V. (2018). *Measuring research: What everyone needs to know*. Oxford University Press.

Waltman, L. (2016). A review of the literature on citation impact indicators. *Journal of Informetrics*, 10(2), 365–391.

Waltman, L., & Van Eck, N.J. (2016). The need for contextualized scientometric analysis: An opinion paper. In Rafols, I., Molas-Gallart, J., Castro-Martínez, E., & Woolley, R. (Eds.), *Proceedings of the 21st International Conference on Science and Technology Indicators* (pp. 541–549). Editorial Universitat Politècnica de Valencia.

Wilsdon, J., Allen, L., Belfiore, E., Campbell, P., Curry, S., Hill, S., ... Johnson, B. (2015). *The metric tide: Report of the independent review of the role of metrics in research assessment and management*. doi:10.13140/RG.2.1.4929.1363