Scholarly Assessment Reports

Moed, H. F. (2020). Appropriate Use of Metrics in Research Assessment of Autonomous Academic Institutions. *Scholarly Assessment Reports*, 2(1): 1. DOI: https://doi.org/10.29024/sar.8

RESEARCH

Appropriate Use of Metrics in Research Assessment of Autonomous Academic Institutions

Henk F. Moed
Sapienza University of Rome, IT
henk.moed@uniroma1.it

Policy highlights

- This paper criticizes a "quick-and-dirty" desktop model for the use of metrics in the assessment of academic research performance, and proposes a series of alternatives.
- It considers often used indicators: publication and citation counts, university rankings, journal impact factors, and social media-based metrics.
- It is argued that research output and impact are multi-dimensional concepts; when used to assess individuals and groups, these indicators suffer from severe limitations:
- Metrics for individual researchers suggest a "false precision"; university rankings are semi-objective and semi-multidimensional; informetric evidence of the validity of journal impact measures is thin; and social media-based indicators should at best be used as complementary measures.
- The paper proposes alternatives to the desktop application model: Combine metrics and expert knowledge; assess research groups rather than individuals; use indicators to define minimum standards; and use funding formula that reward promising, emerging research groups.
- It proposes a two-level model in which institutions develop their own assessment and funding policies, combining metrics with expert and background knowledge, while at a national level a meta-institutional agency *marginally* tests the institutions' internal assessment *processes*.
- According to this model, an *inappropriate* type of metrics use is when a *meta*-institutional agency is concerned directly with the assessment of individuals or groups *within* an institution.
- The proposed model is *not* politically neutral. A normative assumption is that of the autonomy of academic institutions. The meta-institutional entity acknowledges that it is the primary responsibility of the institutions themselves to conduct quality control.
- Rather than having one meta-national agency defining what is research quality and what is not, and how it should be measured, the proposed model facilitates each institution to define its own quality criteria and internal policy objectives, and to make these public.
- But this *freedom* of institutions is accompanied by a series of *obligations*. As a necessary condition, institutions should conceptualize and implement their internal quality control and funding procedures.
- Although a meta-institutional agency may help to improve an institution's internal processes, a repeatedly negative outcome of a marginal test may have negative consequences for the institution's research funding.

This paper discusses a subject as complex as the assessment of scientific-scholarly research for evaluative purposes. It focuses on the use of informetric or bibliometric indicators in academic research assessment. It proposes a series of analytical distinctions. Moreover, it draws conclusions regarding the validity and usefulness of indicators frequently used in the assessment of individual scholars, scholarly institutions and journals. The paper criticizes a so called desktop application model based upon a set of simplistic, poorly founded assumptions about the potential of indicators and the essence of research evaluation. It proposes a more reflexive, theoretically founded, two-level model for the use of metrics of academic research assessment.

Keywords: research assessment; bibliometrics; citation analysis; altmetrics; world university rankings; journal impact measures; autonomous universities; research funding

1. What this paper is about

This contribution[1] focuses on the complex and controversial use of bibliometric or informetric indicators in the assessment of research performance. It aims to shed light on the application context of these indicators. Its principal objective is to propose an application model for the use of indicators in the assessment of academic research, highlight a series of analytical distinctions and useful building blocks, and contribute in this way to enlightening current experiences and further developing best practices in research assessment.

The paper consists of three sections. Section 2 presents an introduction to the use of indicators in research assessment, following the monograph Applied Evaluative Informetrics published by the author (Moed, 2017).[2] Next, Section 3 draws conclusions on the validity and usefulness of the most common indicators of the performance of four principal units of assessment: individual researchers, research groups, research institutions and scientific-scholarly journals. The following indicators are considered:

- Publication and citation counts;
- A university's position in university rankings;
- Journal impact factors;
- Altmetrics and indicators based on full text downloads.

Section 4 proposes an application model for the use of indicators in academic research assessment. From the critical discussion in Section 3, conclusions are drawn on the way in which these measures could be used properly, and on how they should better *not* be used. First, a so called desktop or "quick- and-dirty" application model is discussed, and alternative elements are proposed. Next, a two-level model is sketched for the use of indicators in academic assessment and policy. It provides a framework in which metrics could be used in an appropriate manner. The statements made in this section cross the border of the domain of informetrics and enter that of the evaluation and policy. Finally, important points for further consideration are discussed in Section 5.

2. About applied evaluative informetrics

Research output and impact are multi-dimensional

A first notion is that of the *multi-dimensional* nature of scientific-scholarly research output and impact. **Table 1** presents an overview of major types of research output and impact. It does not claim to be exhaustive, but it highlights a series of key distinctions, particularly between publication-based and non-publication-based outputs, and between scientific-scholarly, educational, technological, economic, social and cultural impacts. Moreover, it gives for each category examples of research outputs. The multi-dimensionality of research performance is also reflected in the list of bibliometric or informetric indicators presented in **Table 2**. This table does not merely include citation and other bibliometric indicators, but also patent-, usage-, research data-, reputation-, network- and infrastructure-based metrics, as well as altmetrics and econometric and webometric indicators.

Distinction between policy, evaluation, analytics and data collection

Figure 1 distinguishes 4 levels of intellectual activity related to research assessment: *policy*, formulating and solving a policy issue; *evaluation*, in which an evaluative framework is specified; *analytics*, analysing empirical data; and relevant *data collection*. These levels are further explained in **Table 3**. Informetrics and bibliometrics deal with the levels of *data*

Table 1: Multi-dimensional output and impact.

Impact dimension	Publication based	Non-publication based
Scientific-scholarly	Scientific journal paper; book chapter; scholarly monograph; conference paper; editorial; review	Research dataset; software, tool, instrument; video of experiment; registered intellectual rights
Educational	Teaching course book; syllabus; text- or handbook	Online course; students completed; degrees attained (e.g., doctorates)
Economic or technological	Patent; commissioned research report	Product; process; device; design; image; spin off; registered industrial rights; revenues from commercialization of intellectual property
Social or cultural	Professional guidelines; policy documents; newspaper article; encyclopaedia article; popular book	Interviews; events; performances; exhibits; scientific advisory work; Communication in social media, e.g., blogs, tweets

[1] Full version of a paper entitled "The Application Context of Research Assessment Methodologies" presented by the author during the conferral of a doctorate honoris causa from the Sapienza University of Rome, 5 September 2019, and at the ISSI2019 conference in Rome on the same day.

[2] Table 1 in Section 2 of the current paper is a summary of Tables 3.2 and 3.3 in Moed (2017, Ch. 3.2–3.3, p. 47–49); Table 2 is a summary and an extension of Table 3.5 in Ch. 3.5, p. 51–59. Table 3 is an extension of Table 6.3 in Ch. 6.3 on page 80; Table 4 is based on the text in Ch. 8.2 on page 120; Table 5 in the current paper's Section 4 summarizes the main lines of Ch. 10, pp. 141–152. The use of the term informetrics reflects that the book – as well as the current paper – does not only deal with bibliometric indicators based on publication and citation counts, but also with altmetrics, webometrics, and usage-based metrics derived from a variety of data sources, as well as research input and process indicators.

Table 2: Twelve informetric indicator families in research assessment.

Indicator	Specification; examples
Publication-based indicators	Publication counts by type of publication
Citation-based indicators	Citation impact, visibility; un-citedness
Journal metrics	Journal impact factor
Patent-based indicators	Patents; patent citations
Usage-based indicators	Full text downloads, html views
Altmetrics	Mentions in social media; readership counts
Webometrics	Web presence; Web linkages
Indicators related to research data	Quality and accessibility of research data
Econom(etr)ic, technology-related indicators	Efficiency (output/input); licenses; spin-offs
Reputation-based measures	Prizes, awards
Network-based indicators	Collaboration, migration, cross-disciplinarity
Indicators of research infrastructure	Facilities, scale, sustainability

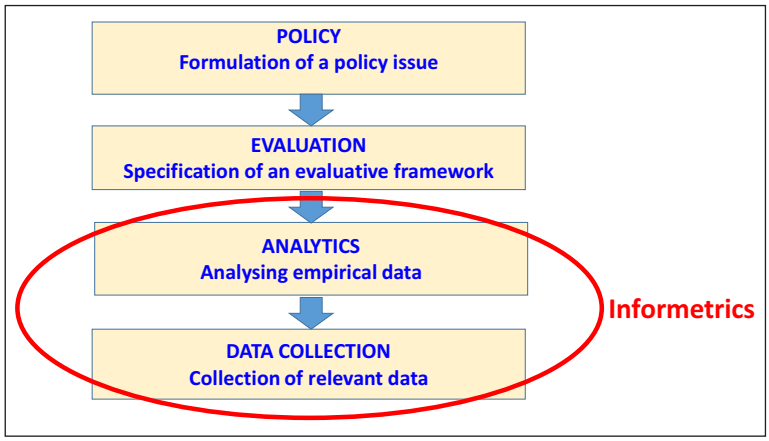

Figure 1: Four levels of intellectual activity in research assessment.

Table 3: Four levels of intellectual activity in research assessment.

Level	Key aspect and issues (examples)	Main outcome
Policy	Desirable objectives; strategies to serve them; Are objectives and strategies fair and aligned with the rules of good governance?	Policy decision based on the outcomes from the evaluation domain
Evaluation	Evaluative framework: what is valuable and how is value to be assessed. What constitutes performance?	A judgment on the basis of an evaluative framework and the empirical evidence collected.
Analytics	Empirical and statistical research: development of new methods; indicator validity; effectiveness of political strategies.	An analytical report as input for the evaluative domain.
Data collection	Creation of databases with data relevant to the analytical framework; Data cleaning; assessment of data quality.	A dataset for the calculation of all indicators specified in the analytical model.

collection and *analytics*. It is true that informetrics has a great potential in the assessment of research and that a large part is still unexplored. But informetrics itself does *not* evaluate. Its application in research assessment requires an evaluative framework. Its validity cannot be established in quantitative-empirical, informetric research. Not in all assessments such an evaluative framework has been developed properly. The lacking of such framework has sometimes been compensated by ad-hoc arguments of evaluators or by using poorly founded assumptions underlying informetric tools.

Informetrics should be "methodologically value free"

A basic notion holds that from what *is* cannot be inferred what *ought to be*. This applies both to political values and objectives at the level of policy, as well as to views of what is valuable in research and what constitutes research quality,

within the context of an evaluative framework. Evaluation criteria and policy objectives are not informetrically demonstrable values. Therefore, informetricians should maintain in their informetric work a neutral position towards such values, and assign a hypothetical status to them. In *this* sense, evaluative informetrics should be "value free".

Value-free does *not* mean that informetric research is not based on values. First of all, like any scientific scholarly activity, it is guided by strict methodological principles. In fact, the requirement of being value-free outlined above is itself a normative statement (a "value"), expressed at the level of methodological rules and principles. Secondly, informetric researchers have political views and preferences as well. Also, as scholars they have views on what constitutes scholarly quality, and may be subjected to assessment processes as well. The principle of being value-free does not mean that informetricians should not have such views, or that they are not allowed to have political preferences. Neither does it mean that in the selection of research subjects, political considerations are not allowed to play a role. But searching for scientific-scholarly truth should not be affected by them. Whether or not informetric observations or conclusions from empirical research are correct or valid, should not depend upon a researcher's political or evaluative views.

How the policy context influences the choice of indicators

The choice of indicators in an assessment study does not only depend upon the type of entity and the performance dimension to be assessed, but also upon the purpose of the assessment and its broader context. **Table 4** lists four key questions that should be addressed in the setup of any assessment study, and gives some typical examples.[3] The relevance of the broader context of an assessment can be illustrated by the following example. If one aims to assess the activities of research groups in scientifically developing countries that have adopted a policy of stimulating their research potential to enter international research and communication networks, it seems defensible to use sensible measures of journal citation impact to assess their researchers' publication strategies. But in case of assessing candidates for vacant research positions in established research universities, it seems to make little sense to rank them according to the average impact factors of the journals in which they have published their papers and select the person with the highest score.

3. Limitations of informetric indicators

As regards the actual use of bibliometric or informetric indicators in research assessment of individuals, research group and institutions, the current author holds the following position.

- Calculating indicators at the level of an individual and claiming they measure by themselves an individual's performance suggests a *false precision*.
- University rankings are influenced by political premisses and objectives.
- The informetric evidence that journal impact factors are good indicators of the quality of the peer review system and of international orientation is weak.
- "Altmetrics should not be used to help evaluate academics for anything important, unless perhaps as complementary measures" (Thelwall, 2014).

Metrics for individual researchers suggest a "false precision"

Performance of an individual and the citation impact of the papers he or she (co-) authored relate to two distinct levels of aggregation. Research is team work; multiple co-authorship is a rule rather than an exception, especially in the natural, life and applied sciences. A crucial issue is how one should assign the citation impact of a team's papers to the performance of an individual working in that team. This issue cannot merely be solved in an informetric way.

The application of fractional counting based on the number of co-authors, or considering the position of an author in the author sequence in the byline of a paper, taking into account corresponding authorship, or using formal statements

Table 4: Key questions to be addressed in the setup of a research assessment study.

Question	Examples
Unit of the assessment?	A country, institution, research group, individual, research field, international network?
Dimension of the research process should be assessed?	Scientific-scholarly impact? Social benefit? Multi-disciplinarity? Participation in networks?
Purpose and objectives of the assessment?	Allocate funding? Improve performance? Increase regional engagement? Budget cuts?
Relevant, general or 'systemic' characteristics of the units?	E.g., a national research community's orientation towards the international research front; or phase of scientific development

[3] For more details, the reader is referred to a report of the Expert Group on the Assessment of University-Based Research, in which the notion of a multi-dimensional research assessment matrix was introduced (AUBR, 2010).

in research papers on author contributions, are per se interesting approaches, but they do not solve the problem of assessing the contribution of an individual to team work.[4] The following examples illustrate some of the severe interpretation problems of publication counts for an individual researcher.

- While in some departments all publications made by doctoral students are as a rule co-authored by their supervisors, in other groups supervisors may be reluctant to feature as an author in articles of their students, and therefore have a low publication output.
- Members of research groups may have different functions. Some members may not conduct "pure" research activities, but nevertheless carry out essential management or fund-raising tasks that are essential for a group's research performance. They may have low publication counts.
- In institutions with oppressive working relations among colleagues, a senior member may force his or her subordinates to become co-author of their papers.

University rankings are semi-objective and semi-multidimensional

A critical analysis of "world" university rankings shows that each ranking system has its own orientation or 'profile', and that there is no 'perfect' or 'objective' ranking system. Their geographical coverage, rating methods, selection of indicators and indicator normalizations, have an immediate effect upon the ranking positions of given institutions.[5] Current ranking systems are still one-dimensional in the sense that they provide finalized, seemingly unrelated indicator values in parallel, rather than offer a dataset and tools to observe patterns in multi-faceted data. The following examples illustrate some of the interpretation problems of "world" university rankings.

- Using a "normalized" indicator that corrects for differences in citation impact across geographical regions may cause "top" universities to be more evenly distributed among regions. A methodological decision to use such indicator boosts up the position of more regionally oriented institutions in a world ranking.
- Both research productivity and graduation productivity are important, mutually dependent aspects of institutional performance. If an institution's number of publications per academic staff increases over time, the number of graduates per staff may decline and vice versa. Considering merely one of these two aspects may easily lead to misinterpretations of an institution's performance (e.g. Bruni et al., 2019).

Informetric evidence of the validity of journal impact measures is thin

The use of journal impact factors (JIFs) and related citation based indicators of journal impact[6] seems to be based on the following two assumptions:

- JIFs are good measures of the quality of a journal's manuscript peer review process.
- JIFs are good measures of the international orientation of a journal's authors and readers.

The current author holds the position that the informetric evidence in support of these assumptions is rather weak. Although several validation studies conducted in the past have reported for selected subject fields a positive (rank-) correlation between journal impact factors and peer opinions on the status or quality of journals, there is *no direct* empirical evidence supporting these two assumptions. To test the *first* assumption, it would be necessary to correlate journal impact measures with citation-independent indicators of the quality of a journal's peer review process. To develop

[4] Fractional counting means for instance that if a paper is published by n authors, it is assigned for a portion 1/n to each author. But more sophisticated fractional counting schemes have been proposed as well, for instance, assigning a fraction 1/2 to the first author of a paper, 1/4 to the last author, and dividing the remaining quarter equally among the other co-authors. As a response to the difficulties with defining authorship in science, several scientific publishers implemented a contributorship model, according to which authors specify their precise role in the research described in a paper, using a classification pre-defined by the publisher. See for instance the website of the Council of Science Editors for more information (www.councilscienceeditors.org). The current author does not see why this feature would actually solve the authorship problems identified by this Council. If there is among some researchers a certain tendency to include authors who do not deserve it, there is little reason to assume that the same tendency would not affect the specification of author contributions as well.

[5] Indicator normalizations are statistical tools to correct for particular "disturbing factors" or biases. A typical example is a citation impact indicator that corrects for differences in citation practices among scientific subfields, by dividing the citation rate of papers published by a group or journal by the world citation average in the disciplines covered by that group or journal. In this way groups or journals in mathematics may have normalized scores similar to those of molecular biologists, while citation levels in the latter discipline are much higher than in the former. For an extensive review on field normalization, see Waltman and Van Eck, 2019. In a similar way, one can normalize the citation impact for differences in impact across geographical regions by dividing the average citation rate of a university by the average in the world region in which that university is located.

[6] The journal impact factor calculated and published by Clarivate Analytics (formerly Thomson Reuters and Institute for Scientific Information) in the Journal Citation Reports (JCR) is defined as the average citation rate in a particular year of articles published in the two preceding years. The San Francisco Declaration on Research Assessment (DORA, 2009) stated that the use of journal-based metrics must be eliminated in evaluation of individuals, and must be greatly reduced in journal promotion (Van Noorden, 2013). Despite this critique, journal impact factors and related measures are still used, not only by librarians, publishers and authors to evaluate scientific journals, but also by research managers and evaluators in the assessment of individual researchers. See for instance Gunashekar, Wooding, & Guthrie (2017), p. 1819 (Table 2), and especially note 14 at bottom of p. 1824. For an extensive review on journal citation measures, the reader is referred to Larivière & Sugimoto, 2019.

such indicators, one could apply computational-linguistic techniques to analyse the full text of referee reports on submitted journal manuscripts, and explore the construction of indicators of reviewers' thoroughness and impartiality (Moed, 2017, p. 161–164). Testing the *second* assumption, a recent study found that the statistical relationship between a journal's index of national orientation and its impact factor is not linear but reverse U-shaped: as a journal's national orientation increases, its impact factor first tends to increase, followed by a decline (Moed et al., 2020).

Altmetrics should be used with great caution as complementary measures

Altmetrics relates to different types of data sources with different functions.[7] Mentions in social media may reveal impact upon non-scholarly audiences, and provide tools to link scientific expertise to the interest of the wider public, and to societal needs. Indicators derived from scholarly reference managers such as Mendeley and ResearchGate give insight into the publication "kitchen" in which new articles are being prepared, and are therefore potentially faster predictors of emerging scholarly trends than citations are. "Usage" data on downloads of an electronic publication in html or full text format enable researchers to assess the effectiveness of their communication strategies, and may reveal attention of scholarly audiences from other research scientific domains or of non-scholarly audiences.

Wouters, Zahedi & Costas (2019) conclude that social media metrics are new evaluation tools, reflecting social presence, reception and engagement, and "communities of attention around scholarly objects". But the application of such metrics and related in research assessment of individuals has severe limitations as well. Below follow some of these limitations.

- Although social media metrics may reflect attention of a wider public, they should not be used to measure scientific-scholarly impact. Their numbers can to some extent be manipulated, "since social websites tend to have no quality control and no formal process to link users to offline identities" (Thelwall, 2014).
- Visibility of researchers in social media and reference managers strongly depends upon the extent to which they themselves decide to actively participate in such media.
- Readership counts in scholarly reference managers depend upon readers' cognitive and professional background, and need not necessarily be representative for a wider scientific-scholarly audience (Mohammadi & Thelwall, 2019).
- Downloaded articles may be selected according to their face value rather than their value perceived after reflection. Also, there is an incomplete usage data availability and a lack of standardisation across providers, and counts can to some extent be manipulated (Kurtz & Bollen, 2010).

4. A model for the use of metrics in evaluation and policy

Alternatives to a "quick-and-dirty" desktop application model

Bibliometric or informetric indicators are often applied in what is termed a "desktop assessment" model, that is based on a series of simplistic, seemingly plausible, but, upon reflection, questionable assumptions (Katz & Hicks, 1997). They are summarized in the middle column of **Table 5**. In the third column of this table, alternative approaches are high-

Table 5: The assumptions of the desktop assessment model and alternative approaches.

Aspect	Desktop model assumptions	Alternative approaches
Availability	Widely available indicators should be used (publication counts, journal impact factors, h-indices)	Use tailor-made indicators appropriate for the *application context*
Validity	Indicators measure well what they are supposed to measure; no confirmation from other sources is needed	Indicators should be *combined* with *expert knowledge* and measure *pre-conditions* of performance rather than performance itself
Evaluative significance	The aspects measured by the indicators constitute appropriate evaluation criteria	An *independent evaluative* framework is needed
Unit of assessment	Evaluate individual researchers	Do *not* separate a researcher from his/her institutional context: Focus on research *teams*
Ordering principle	The higher the score, the better the performance	Use indicators to define *minimum standards* rather than identifying the top of a ranking
Policy decision criteria	The best overall performer receives the largest support	Fund institutions according to the number of *emerging* research groups

[7] The concept of "Altmetrics" is introduced in an Altmetrics Manifesto published on the Web in October 2010 (Priem et al., 2010). This manifesto gives an overview of the increasingly important role of social media such as Twitter and Facebook, online reference managers such as Mendeley, ResearchGate and Zotero, scholarly blogs and online repositories in scientific scholarly communication. It underlines that the activities in these online tools can be tracked: "This diverse group of activities forms a composite trace of impact far richer than any available before. We call the elements of this trace altmetrics". A good overview of the challenges of altmetrics is given by Haustein (2016). The Springer Handbook on Science and Technology Indicators includes chapters on social media metrics (Wouters, Zahedi and Costas, 2019) and readership-based measures (Mohammadi and Thelwall). An overview paper about the potential and limits of altmetrics will be published soon in Scholarly Assessment Reports in 2020.

lighted, that are further developed in Moed (2017, p. 141–152). A full discussion goes beyond the scope of the current paper. But several of the proposed elements can be positioned in a general application model schematized in **Figure 2**.

A two-level model based on institutions' autonomy

With respect to the application of informetric data and indicators, **Figure 2** distinguishes an informetric (statistical, analytical), evaluative and political component, in line with the distinctions in levels of intellectual activity in research assessment summarized in **Figure 1** and **Table 3** in Section 2. Another important element in **Figure 2** is a distinction between an institutional and a meta-institutional level. The *institutional* level relates to *internal* assessment and funding processes *within* a university. The meta-institutional level deals with the national academic system *as a whole*, and with the distribution of funds among institutions.[8] Typical examples of agencies with meta-institutional tasks and responsibilities are a Ministry of Higher Education or Research, or a National Accreditation Organization.

Figure 2 shows two application lines of informetric indicators in research assessment and policy. The *bottom* application line relates to the use of indicators *within* an institution. As outlined in Section 3, making proper evaluations and informed decisions about individuals or groups require background knowledge about these units of assessment, their fields, and their institutional context. This knowledge tends to be unavailable to external meta-institutional entities operating at a large distance from an institution. It is within institutions that indicators are combined with background and expert knowledge. They provide an input into internal evaluation and funding processes. These processes may relate to hiring and promotion of individuals, but also to the distribution among research groups of research funds drawn from a special research budget that institutions can use freely to conduct a proper internal research policy.[9]

The *upper* application line relates to the meta-institutional (e.g., national) level. To the extent that agencies active at this level apply indicators about institutions, the model proposes such measures to be statistical aggregates, calculated for an institution in its entirety. A possible application uses such aggregate indicators to calculate parameters in a funding formula used to distribute funds across institutions.[10] A second type of use on the supra-institutional level concerns the assessment of *evaluation processes* carried out within the institution. In such an assessment, a meta-national agency

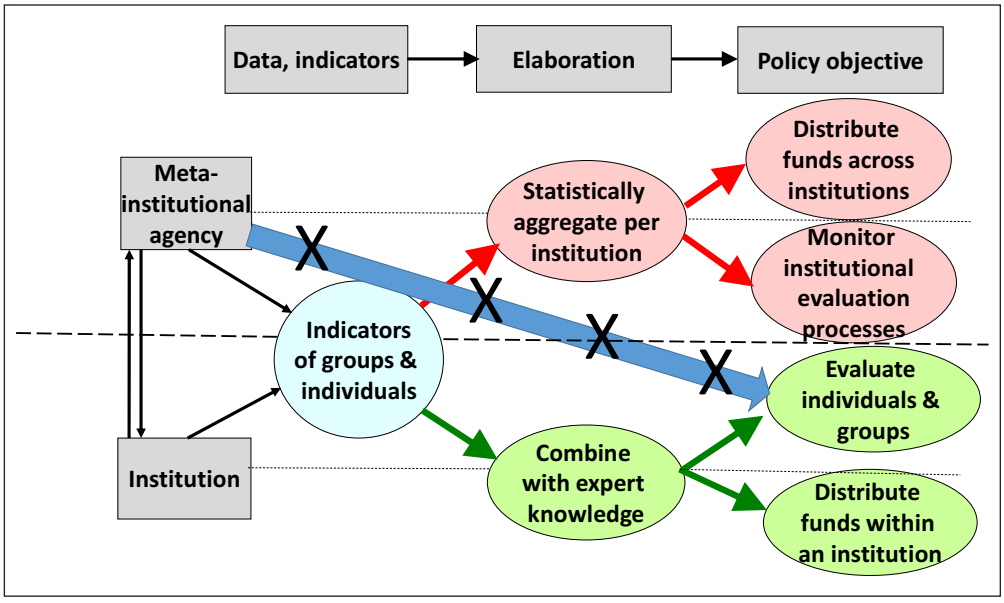

Figure 2: A multi-level application model of informetric indicators of research performance. While institutions in their internal assessment and funding processes combine indicators and expert knowledge to evaluate individuals and groups, a meta-university agency aggregates indicators at the level of an institution, and assesses the evaluation and funding *processes* inside an institution. The crossed arrow indicates an inappropriate type of use of indicators in which a meta-institutional agency is concerned directly with assessment of individuals or groups at the institutional level.

[8] The use of the term "as a whole" does not preclude the possibility to conduct assessments on a discipline-by-discipline basis, evaluating, for instance, all research in humanities, or all research in biomedicine.

[9] In the Flemish academic system, universities are free to spend a part of their basic funding (Special Research Funds, in Dutch: Bijzonder Onderzoeksfonds (BOF)) applying their own criteria. More information on this system is given by Marc Luwel in a forthcoming contribution to this journal.

[10] Nowadays many countries have performance-based funding processes, distributing (a part of) funds among universities. See for instance OECD (2010) and Zacharewicz et al. (2018). In several countries, bibliometric indicators constitute one of the parameters in a funding formula. An important issue is how large the role of metrics in national research assessment exercises can or should be. Harzing (2018) analysed the outcomes of the Research Excellence Framework (REF) in the UK and "was able to create a citation-based ranking of British universities that correlates with the REF power ranking at 0.97", suggesting that the effort and costs of a national exercise can be substantially reduced by giving a more prominent role to metrics, thus creating a financial space for tailor-made, qualitative assessments. In the current paper this type of use of indicators is not discussed further.

marginally tests the procedures along which an institution carries out its internal assessment and funding policies. In these tests, indicators may constitute one of the sources of information, but it is *not* the position of an individual researcher or group within an institution that is at stake, but the defensibility and the effectiveness of the overall process of quality control of the institution as a whole.

The genuine challenge in the responsible and fair use of informetric indicators of research performance in an academic environment does not primarily lie in the further sophistication of indicators, but in the ability to establish external, independent and knowledgeable entities who monitor the evaluation processes within institutions, acknowledging that it is the primary responsibility of the institutions themselves to conduct quality control.

5. Important points to consider

Key notion of the proposed model

The proposed model sketches the main lines of two ways in which bibliometric methods could be used in academic research assessment. A key notion underlying the model is that a meta-institutional agency could not possibly carry out evaluation at the level of groups and individual researchers without recourse to a set of single-measure indicators that would ultimately be used in mechanical ways; what is needed at this level is expertise, time, and sufficient resources to conduct an assessment that is qualitative in nature, and possibly informed by quantitative indicators.

Application contexts and conditions

The current paper does not present a full discussion of the conditions under which the model can be applied, and the contexts in which it fits best. It is primarily the research policy domain that deals with proper conditions and contexts. Large differences exist in academic research funding models among countries. For instance, in the US research funding tends to be dispersed based on competition among individual researchers and not principally through block grants to institutions that can be found in many continental European countries. It seems more recommendable to implement the proposed model in the latter group of countries than in the first, even though certain elements could be useful in both groups. Of course, the challenge in implementing the model is attaining a relationship of trust between the research funder and the academic research community. The proposed model will not work if there is a fundamental lack of trust between these two parties, but it can further increase trust once it is implemented.

Selection of bibliometric/informetric indicators

The paper does not specify in detail which bibliometric/informetric methodologies and indicators are the most appropriate. As argued in Section 2, the choice of informetric indicators depends upon what is being assessed, and which criteria are to be used, and also upon the purpose of the assessment and its broader context. A set of indicators useful in one country may be less appropriate in another. For instance, in some countries there is a growing liaison between academic and industry research, driven through government funding. In other countries, hosting large publicly funded institutes of applied research, government priorities in the funding of academic research may have a somewhat different focus. An Expert Group on the Assessment of University-Based Research Expert Group (AUBR, 2010) presented a matrix listing indicators for a series of policy objectives. It is up to the informetric research community to update and further develop such a matrix.

Separate quality assessment processes from funding decisions

The current author holds the position that the supra-institutional agency that is responsible for the assessment of internal quality processes should give a qualitative judgement, even though quantitative indicators may help the agency to form and motivate its judgement. Also, metrics can help the agency to ask relevant, critical questions to institutional managers about their internal processes. These types of use are *not* aimed at reaching a quantitative outcome that can be uploaded directly into a funding formula. In terms of a distinction between *formative* and *summative* evaluation,[11] it can be said they have a more formative nature, whereas those linked to funding are more summative in nature. The task to assess institutions' quality processes and the calculation of funding formula should *not necessarily* be carried out by one and the same agency. It is a defensible practice to separate the formation of a judgement on the quality in internal assessment processes from the task of defining funding formula and making decisions to distribute funds among institutions.

The proposed model is not politically neutral

The model proposed is based on the insight that internal evaluations within institutions, combining indicators and expert knowledge and including views of external experts, constitute more favourable conditions for a proper use of the indicators than purely informetric-statistical use by external, meta-institutional agencies do. The model is *not*

[11] Research assessment has both a formative and a summative function. In summative evaluation the focus is on the outcome of a process or program, such as a final judgement or a vote. Formative evaluation assesses the unit's development at a particular time, and primarily aims to improve its performance. Several researchers have rightly underlined the value of formative evaluations, and consider the use of bibliometric or altmetric indicators predominantly as formative tools. But especially in the context of research funding, research assessment has a summative aspect as well. This is true for a supra-institutional agency assessing internal evaluation and funding processes, but also of the intra-institutional assessment of individuals and groups. But this does not imply that assessment outcomes are necessarily expressed in numbers.

politically neutral. A "normative" assumption is that of the autonomy of academic institutions. A meta-institutional entity acknowledges that it is the primary responsibility of the institutions themselves to conduct quality control. It stimulates institutions to profile themselves on the basis on how they define and implement a notion as complex as academic research quality. Rather than having one single meta-national agency defining what is research quality and what is not, and how it should be measured, the proposed model lets each institution define and operationalize its own quality criteria and internal policy objectives, and make these public.

Institutions have autonomy but also obligations

But this *freedom* is accompanied by a series of *obligations*. As illustrated above, the meta-institutional agency marginally assesses the process and its overall outcomes, possibly using appropriate indicators. In case of serious doubts, an audit by independent experts may be in place, focusing on the intra-institutional processes, and aiming to their improvement. If the outcomes of the audit do not lead to improvement of intra-institutional processes, this may have negative consequences for the funding of an institution. Therefore, as a necessary condition, institutions should make further steps in the conceptualization and implementation of their internal quality control and funding procedures.

The interaction between the supra-institutional and the institutional level must be considered

An important issue is the optimal relationship between institutional and meta-institutional or national assessment processes. Is it defensible that the two levels apply the same indicators? How to minimize possible negative effects of meta-institutional assessment procedures upon intra-institutional behaviour and evaluation? These questions concern the effects that the application of indicators at the meta-institutional level may have upon the assessment and research practices within the institution.

The issue at stake is not whether the application of metrics lead to changes in these intra-institutional practices, but whether or not such changes reflect a genuine enhancement of the performance of an institution. If there is solid evidence that certain quantitative criteria applied by a meta-institutional agency induce at the intra-institutional level a systematic, strategic behaviour aimed to obtain a high score rather than a substantive quality enhancement, the system is counter-productive. The same conclusion holds when research groups within institutions, aware as they are of the funding formula applied at the meta-institutional level, claim an amount of funding within their institution proportional to the contribution they make to the parameters in the formula.[12]

Institutional self-profiling is crucial but standardisation across institutions is needed as well

On the one hand, institutional self-profiling is essential. The model presupposes that institutions define their own profile and targets. For instance, an institution's disciplinary specialization is an important characteristic; the same is true for its role in regional development. As indicated in **Table 4** in Section 2, the choice of indicators partly depends upon these characteristics. But standardisation is to some extent necessary as well, both in the intra- and in the meta-institutional assessment process. Standardisation may take place on a discipline-by-discipline basis, acknowledging that indicators apt to use on one discipline may be of little validity in another.

Benchmarking, this means comparing an institution with other institutions with similar profiles, can be expected to play an important role in the assessment of an institution's quality and funding processes. In order to treat each institution in the same manner, the selection of appropriate benchmarks is crucial. Informetric tools can be useful to suggest potential benchmarks. These tools generate indicators, not of the standard publication output or citation impact, but, for instance, of the degree of similarity among the orientations of a wider set of institutions. This is another example of how bibliometric-informetric indicators can be properly used in academic research assessment.

Acknowledgements

The author wishes to thank Dr Marc Luwel, former director of the Netherlands-Flemish Accreditation Organization in The Hague, for stimulating discussions and for his useful comments on an earlier version of this paper. He is also grateful to the two reviewers for their valuable comments.

Competing Interests

The author has no competing interests to declare.

References

AUBR. (2010). Assessment of University-Based Research Expert Group (AUBR). Assessing Europe's University-Based Research. *K1-NA-24187-EN-N, European Commission*, Brussels (pp. 151). http://ec.europa.eu/research/era/docs/en/areas-of-actions-universities-assessing-europeuniversity-based-research-2010-en.pdf

[12] Empirical research on the effects of the use of indicators in performance-based funding shed light on whether these practices have played an important role. More and more studies are being published on this subject. See for instance a thorough review by De Rijcke et al., 2016. The challenge is to give systematic, well documented, scholarly accounts of these effects, thus reaching beyond the level of personal experiences and impressions, genuine and valid as they may be.

Bruni, R., Catalano, G., Daraio, C., Gregori, M., & **Moed, H. F.** (2019). Characterizing the Heterogeneity of European Higher Education Institutions Combining Cluster and Efficiency Analyses. In: *Proceedings of ISSI2019*, Rome, 2–5 September 2019.

De Rijcke, S., Wouters, P. F., Rushforth, A. D., Franssen, T. P., & **Hammarfelt, B.** (2016). Evaluation practices and effects of indicator use—a literature review. *Research Evaluation*, *25*(2), 161–169. DOI: https://doi.org/10.1093/reseval/rvv038

DORA. (2009). *San Fransisco Declaration on Research Assessment*. Available at http://www.ascb.org/dora/

Gunashekar, S., Wooding, S., & **Guthrie, S.** (2017). How do NIHR peer review panels use bibliometric information to support their decisions? *Scientometrics*, *112*, 1813–1835. DOI: https://doi.org/10.1007/s11192-017-2417-8

Harzing, A. W. (2018). Running the REF on a rainy Sunday afternoon: Can we exchange peer review for metrics? *Leiden: STI 2018 Conference Proceedings* (pp. 339–345). https://openaccess.leidenuniv.nl/handle/1887/65202

Haustein, S. (2016). Grand challenges in altmetrics: heterogeneity, data quality and dependencies. *Scientometrics*, *108*, 413–423. DOI: https://doi.org/10.1007/s11192-016-1910-9

Katz, J. S., & **Hicks, D.** (1997). Desktop Scientometrics. *Scientometrics*, *38*, 141–153. DOI: https://doi.org/10.1007/BF02461128

Kurtz, M. J., & **Bollen, J.** (2010). Usage bibliometrics. *Annual review of information science and technology*, *44*, 1–64. DOI: https://doi.org/10.1002/aris.2010.1440440108

Larivière, V., & **Sugimoto, C.** (2019). The Journal Impact Factor: A Brief History, Critique, and Discussion of Adverse Effects. In: W. Glanzel, H. F. Moed, U. Schmoch & M. Thelwall (Eds.), *Springer Handbook of Science and Technology Indicators* (pp. 3–23). Switzerland: Springer Nature. DOI: https://doi.org/10.1007/978-3-030-02511-3_1

Moed, H. F. (2017). *Applied Evaluative Informetrics* (pp. 312). Springer International Publishing. DOI: https://doi.org/10.1007/978-3-319-60522-7

Moed, H. F., de Moya-Anegon, F., Guerrero-Bote, V., & **Lopez-Illescas, C.** (2020). Are nationally oriented journals indexed in Scopus becoming more international? The effect of publication language and access modality. Manuscript submitted for publication.

Mohammadi, E., & **Thelwall, M.** (2019). Readership Data and Research Impact. In: W. Glanzel, H. F. Moed, U. Schmoch & M. Thelwall (Eds.), *Springer Handbook of Science and Technology Indicators* (pp. 761–775). Switzerland: Springer Nature. DOI: https://doi.org/10.1007/978-3-030-02511-3_29

OECD. (2010). Performance-based funding for public research in tertiary education institutions: workshop proceedings. OECD Publishing. DOI: https://doi.org/10.1787/9789264094611-en

Priem, J., Taraborelli, D., Groth, P., & **Neylon, C.** (2010). Altmetrics: A Manifesto. Available at http://altmetrics.org/manifesto/

Thelwall, M. (2014). A brief history of altmetrics. *Research Trends*, issue 37 (Special issue on altmetrics, June). Available at http://www.researchtrends.com/issue-37-june-2014/a-brief-history-of-altmetrics

Van Noorden, R. (2013). Scientists Join Journal Editors to Fight Impact-Factor Abuse. *Nature News Blog.* 16 May 2013. Available at http://blogs.nature.com/news/2013/05/scientists-join-journaleditors-to-fight-impact-factor-abuse.html

Waltman, L., & **Van Eck, N. J.** (2019). Field normalization of scientometric indicators. In: W. Glanzel, H. F. Moed, U. Schmoch & M. Thelwall (Eds.), *Springer Handbook of Science and Technology Indicators* (pp. 281–300). Switzerland: Springer Nature. DOI: https://doi.org/10.1007/978-3-030-02511-3_11

Wouters, P., Zahedi, Z., & **Costas, R.** (2019). Social Media Metrics for New Research Evaluation. In: W. Glanzel, H. F. Moed, U. Schmoch & M. Thelwall (Eds.), *Springer Handbook of Science and Technology Indicators* (pp. 687–714). Switzerland: Springer Nature. DOI: https://doi.org/10.1007/978-3-030-02511-3_26

Zacharewicz, T., Lepori, B., Reale, E., & **Jonkers, K.** (2018). Performance-based research funding in EU Member States—a comparative assessment. In *Science and Public Policy* (pp. 1–11). DOI: https://doi.org/10.1093/scipol/scy041

How to cite this article: Moed, H. F. (2020). Appropriate Use of Metrics in Research Assessment of Autonomous Academic Institutions. *Scholarly Assessment Reports*, 2(1): 1. DOI: https://doi.org/10.29024/sar.8

Submitted: 06 November 2019 **Accepted:** 02 January 2020 **Published:** 22 January 2020

Scholarly Assessment Reports is a peer-reviewed open access journal published by Levy Library Press. **OPEN ACCESS**