Guidance on the responsible use of quantitative indicators in research assessment

Table of contents

Produced by the DORA Research Assessment Metrics Task Force:

Ginny Barbour
Rachel Bruce
Stephen Curry
Bodo Stern
Stuart King
Rebecca Lawrence

Please cite this document as:

DORA. 2024. Guidance on the responsible use of quantitative indicators in research assessment. http://doi.org/10.5281/zenodo.10979644

For more information, contact Zen Faulkes, DORA Program Director.

Introduction

Research assessment is an important and challenging task and many institutions work hard to grapple with its complexities. Nevertheless, the tendency to fall back on quantitative indicators (or metrics[1]) that are often assumed to provide a measure of objectivity remains widespread. While indicators have great utility in the fields of bibliometrics and scientometrics (e.g., tracking the growth or decline of different subfields), they are inherently reductive so their use in the assessment of individual researchers or research projects requires careful contextualization.

The Declaration on Research Assessment (DORA) is best known for being critical of the misuse of the Journal Impact Factor (JIF) in research evaluation. As a result, DORA is often asked for its views on other indicators. In this briefing note we therefore aim to explain how the principles underlying DORA apply to other quantitative indicators that are sometimes used in the evaluation of research and researchers.

[1] While the term "metric" suggests a quantity that is being measured directly, "indicator" better reflects the fact that quantities being used in research assessment are more often indirect proxies for quality. We therefore use "indicator" throughout this briefing document.

A close reading of the Declaration on Research Assessment reveals an approach to the use of quantitative information that is based on five simple principles:

Be clear

What is your rationale for using particular quantitative indicators in your research or researcher assessments? Is it grounded in good evidence?

Be transparent

Ideally, rules for the use of quantitative indicators in research assessment should be developed in dialogue with your research community.[2] They should be published so that those being evaluated understand your criteria. Make sure also that reviewers are fully aware of your approach to using quantitative information in assessment.

Be specific

How well does the indicator refer to the qualities of the person or the piece of work being assessed? Be mindful of aggregate metrics (e.g., JIF, *h*-index), which conceal large variations in performance, and of composite indicators (e.g., scores in university league tables, altmetrics), which are made up of arbitrarily weighted scores for very different attributes and activities and are therefore difficult to interpret meaningfully.

Be contextual

How will you take account of the proxy and reductive nature inherent in any indicator? (E.g., citations are not a direct measure of quality; the *h*-index takes no account of age, discipline, or career breaks.)

Be fair

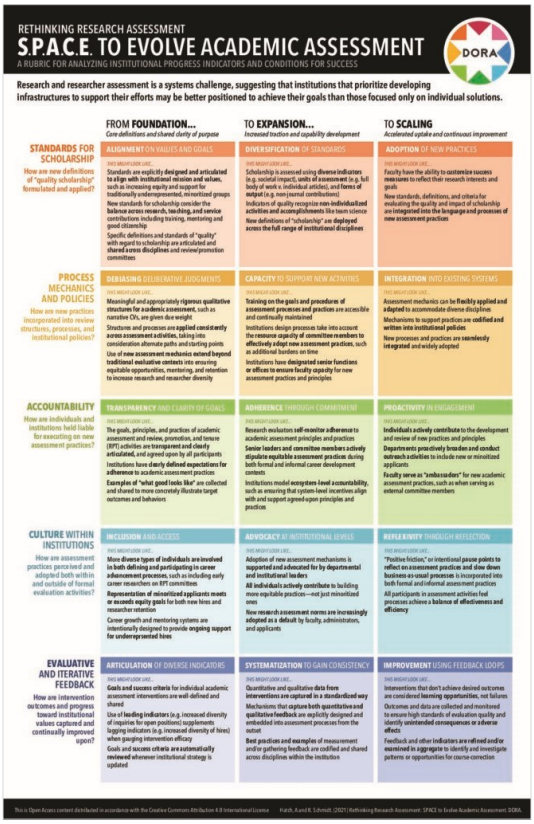
How will you avoid biases inherent in quantitative indicators? Though it is often assumed that bibliometric indicators are "objective," decisions to publish a paper or to cite it are choices that can reflect structural and personal biases. Decision makers need to be proactive and transparent in efforts to mitigate the impact of these biases in research assessment — and the same obviously applies to the qualitative aspects of assessment.

[2] Ideally also, any indicator used should be based on open data and algorithms so that anyone being evaluated can verify how it is calculated but many commonly used "off the shelf" indicators still rely on closed data.

Below we explain how these principles apply to some of the more commonly used indicators. The list cannot be exhaustive, but we hope these examples will show how the principles could be applied in practice to any quantitative indicator. Giving undue weight to just one or two indicators is unlikely to provide a properly informed or balanced evaluation. Best practice is to co-create research assessment processes with your organizational community and to start by agreeing on the values, outcomes, and behaviors that will set the benchmarks for your assessment. The INORMS SCOPE framework or DORA's SPACE rubric are useful tools for this purpose.

The SCOPE framework is created by INORMS.

The SPACE Rubric is available in the DORA Resource Library.



Journal Impact Factor (and other journal metrics)

The Journal Impact Factor (JIF) is an indicator that can essentially be defined as the annual average number of citations to papers in any given journal in the two preceding years. (The actual calculation is more opaque than this.)

Criticisms of the JIF are laid out in some detail in the DORA declaration and elsewhere, but the critical issue for research assessment is that claims that the JIF is a signifier of the value or quality of an individual paper are not supported by a close examination of the evidence. The JIF is a measure of what might be termed the "average citation performance" of papers in a particular journal but, aside from its many technical shortcomings, it gives no indication of the variation in the distribution of citations, which typically range over 2-3 orders of magnitude, from which it is calculated. Although it is tempting to rely on the law of averages and conclude that a paper from a high JIF journal is likely to be better than one from a low JIF journal, the evidence shows that JIFs are poor predictors of citation performance of individual articles. Further, it is also often stated that the quality of peer review is higher at high JIF journals, but we know of no good evidence to support this.

So, when judging individual publications or their authors, one has to look closer. The individual citation performance of the paper can provide some insight but, as discussed below, needs to be considered in context. Assessment of the content is also critical, as is knowing the particular contribution of any author listing it in their CV. Narrative CVs are emerging as a useful tool for capturing this more qualitative information in concise and comparable forms.

The reservations noted here regarding the use of the journal impact factor apply equally to other journal-based indicators, e.g., the Citescore, the Eigenfactor Score, and the Source Normalized Impact per Paper (SNIP).

Citations

The citation count of an article is defined as the number of times it is included in the reference list of other articles or books. At first glance, using article citations in researcher assessment is an improvement over journal-based indicators like the journal impact factor, because citations offer information at the relevant level of granularity, the individual research article. However, as with any quantitative indicator, citations provide a limited view of researcher performance.

Citation performance is a lagging indicator that takes time, often years, to turn into a robust signal. It is therefore not well suited to evaluate recent scholarship or to compare researchers at different career stages, or in different disciplines.

Any use of citations in research assessment should also bear in mind other limitations. Bibliometricians acknowledge that citations reflect the influence of a research article, but this can differ in important ways from what evaluators may really want to determine: the quality and significance of research findings. Citation patterns can be skewed by author and journal reputations; e.g., author status can lead to citation bias, with prominent researchers attracting more citations for similar work than less well-known researchers, a phenomenon long known as the Matthew effect. Likewise, citations of identical editorials published in multiple journals correlate with the Journal Impact Factor. Numbers of citations are also impacted by the variable publication volumes of different disciplines; citations should therefore not be used to compare researchers in dissimilar fields. Differences in citation patterns that disfavor women are also well documented and should be taken into account when considering citations for researcher assessment. Moreover, since citation data do not indicate whether articles are cited for positive or negative reasons, they cannot be used to indicate research quality without additional supporting information; work to develop Citation Typing Ontology may help to resolve this issue in the future.

For all these reasons, citation data cannot replace the critical judgment of experts and should be used with caution in researcher assessment. An indicator that reflects to what extent subsequent research builds upon a reported discovery would be a significant improvement on current citation-based metrics, since re-use of research findings signifies rigor and significance, two key features of high-quality research.

h-index

The *h*-index for individual authors is defined as the number of their papers that have been cited at least *h* times; an author with a *h*-index of 10, for example, has ten papers, each with at least ten citations.

The *h*-index is commonly used by institutions and individuals to compare researchers or to monitor their "performance" over time. However, it is difficult to interpret meaningfully, not least because it can give inconsistent and counterintuitive readings of researcher impact. Moreover, the value of the *h*-index depends on the database used to derive it (e.g., Web of Science, Scopus, Google Scholar) and can be manipulated by gaming.

As a reductive aggregate indicator, the *h*-index also lacks crucial contextual information that should be included in responsible research assessment. For example, the *h*-index will usually be higher for researchers at later career stages, or who have not taken career breaks, or who work in disciplines that attract higher citation rates (e.g., medical sciences as compared with mathematics or humanities); nor does it take account of the nature of the author's contributions to each of their papers. In disciplines that rely increasingly on interdisciplinary, collaborative approaches, the *h*-index may thus reflect participation in large teams rather than individual contributions.

Any organization making use of the *h*-index in research assessment should be able to explain how it provides a meaningful insight into individual research performance, and how account is taken of individual circumstances (e.g., academic age, career breaks, scholarly discipline).

Field-normalized citation indicators

Commonly used field-normalized citation indicators such as Field Weighted Citation Impact (FWCI) or Relative Citation Ratio (RCR) represent attempts to correct for the citation variability arising from differences between fields, types, and ages of publications. FWCI is calculated typically for a collection of publications as the average ratio obtained by dividing the average number of citations accrued per paper in the collection by the average expected for papers of the same type (e.g., primary research articles) and year of publication that are in the same field. It is therefore an indicator of the relative citation performance of a body of research work. For instance, an FWCI of 2 means the research has twice its expected number of citations for papers in a given subject area.

Caution is necessary when using indicators such as FWCI, not only because of the difficulty in defining which papers belong in which fields (which affects the denominator in the calculation), but also because of the variation inherent in the numbers of citations attracted by the papers making up any given body of work (similar to the skewed citation distributions characteristic of any given journal). Analysis shows that in datasets comprising only a few tens or hundreds of papers, the average FWCI is less reliable because of the impact of highly cited outliers. The FWCI should therefore only be applied to large datasets, typically comprising thousands of papers, e.g., the aggregate output of a large department. Even then, the variability associated with differences in sample size means it should not be reported beyond a single decimal place. It is not suitable for evaluating individual researchers because it is unreliable at the scale of a typical bibliography and can fluctuate significantly over time.

The RCR is an article-level indicator that correlates strongly with the FWCI across numerous subject areas, and has elicited similar concerns about its reliability and suitability for researcher assessment.

Altmetrics

Altmetrics, a generalization of the term "alternative metrics," attempt to capture the amount of attention a research output has received in non-academic outlets (e.g., organizational reports, social media). Types of activities captured within the metric score vary enormously, from those more focused on public engagement (e.g., tweets and reposts, Facebook mentions, newspaper or YouTube coverage), through to researcher engagement (e.g., patents, numbers of post-publication peer reviews, inclusion in research highlight platforms), and even inclusion in policy documents. Different types of altmetric scores, which can be calculated for articles, books, data sets, presentations, and more, can be obtained from a range of commercial providers, including Altmetric, ImpactStory, Plum Analytics, and Overton.

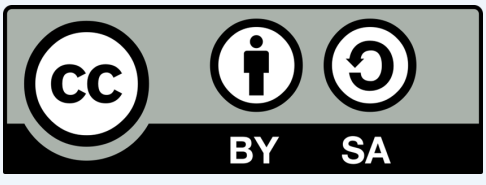
Altmetric information is often presented as a composite score, which represents a weighted measure of all the attention picked up for a research output (i.e., not a raw total of the number of mentions). The weightings used vary for different types of attention and can change as the organizations that produce these scores reassess periodically how best to create the composite figure and as different contributing sources are added or removed over time. It is also important to note that some of the activities included in altmetric scores, especially those associated with social media, are prone to being gamed.

Because of the relatively opaque ways they are calculated, altmetric scores provide little context for the type and purposes of engagement with particular research outputs and are difficult to interpret in terms of broader research impact. They are not in any meaningful sense a measure of research quality.

However, when details of the original mentions and references that contribute to these altmetric scores are provided, these might provide useful information in a more specific context about the levels of attention and reach of a research output (e.g., interest generated among patient advocacy groups). In such circumstances, they may be a useful component of a broader examination of research contributions.

Concluding remarks

The guidance given in this briefing note is neither exhaustive nor comprehensive, but it illustrates how the principles laid out in the DORA declaration can be applied when other metrics are considered for use in assessment of research or researchers. The examples included here refer only to publication-based metrics, but other indicators should be treated in the same way (e.g., see the Metrics Toolkit and the challenges associated with making targets of metrics). For example, grant funding income is often assessed during researcher evaluation since the ability to win competitive funding for ideas is a desirable attribute, but this information should always be contextualized. For example, it is important to recognize that the requirement for funds differs markedly between fields (even within STEM disciplines), that biases still disfavor women and other under-represented groups, and that even the most rigorous funding decisions are attended by uncertainty and are poorly predictive of research productivity.



DORA

6120 Executive Blvd., Rockville, MD, USA

sfdora.org