Contents lists available at ScienceDirect

Journal of Informetrics

Opinion paper

Critical rationalism and the search for standard (field-normalized) indicators in bibliometrics

Lutz Bornmann [a],*, Werner Marx [b]

[a] *Division for Science and Innovation Studies, Administrative Headquarters of the Max Planck Society, Hofgartenstr. 8, 80539, Munich, Germany*
[b] *Max Planck Institute for Solid State Research, Information Service, Heisenbergstrasse 1, 70506, Stuttgart, Germany*

ARTICLE INFO

ABSTRACT

Bibliometrics plays an increasingly important role in research evaluation. However, no gold standard exists for a set of reliable and valid (field-normalized) impact indicators in research evaluation. This opinion paper recommends that bibliometricians develop and analyze these impact indicators against the backdrop of Popper's critical rationalism. The studies critically investigating the indicators should publish the results in such a way that they can be included in meta-analyses. The results of meta-analyses give guidance on which indicators can then be part of a set of indicators used as standard in bibliometrics. The generation and continuous revision of the standard set could be handled by the International Society for Informetrics and Scientometrics (ISSI).

© 2018 Elsevier Ltd. All rights reserved.

1. Introduction

Bibliometrics plays an increasingly important role in research evaluation (Bornmann, 2017). An evaluation report without bibliometrics is currently difficult to imagine. In the application of bibliometrics in evaluative practice, one can differentiate between professional and citizen bibliometrics (Leydesdorff, Wouters, & Bornmann, 2016). Citizen bibliometrics is based on simple indicators, such as the median and total number of citations. These indicators can possibly be used to compare similar units (concerning size and age) in the same or very similar fields. An important disadvantage of these indicators is, however, that they cannot be used in cross-field comparisons of units (e.g. research groups or institutions). The publication and citation cultures are so different in the fields that the median and total number of citations from different fields are not comparable.

In professional bibliometrics, established field-normalized indicators are applied in cross-field comparisons, which take into account that citations occur in different publication and citation cultures (Waltman, 2016). Normalized indicators can be used to assess whether a unit (e.g. an institute) shows above or below average performance in the corresponding fields. Most of the studies in bibliometrics – scientific studies and applied studies used to support research evaluations – are actually cross-field comparisons, because bibliometrics is especially suitable for greater units (e.g. institutes or countries) and these units publish as a rule in different fields. We argue in this opinion paper that bibliometrics should aim to develop standard approaches especially to field-normalization. Only with standard approaches, the results of different studies are comparable, reliable, and interpretable.

---

* Corresponding author.
  E-mail addresses: bornmann@gv.mpg.de (L. Bornmann), w.marx@fkf.mpg.de (W. Marx).

In bibliometrics, however, several variants of normalized indicators exist. For example, mean-based indicators compare the mean impact of the focal papers with the mean impact of papers published in the same field and publication year. Percentile-based indicators assign to every paper in a certain field and publication year a citation percentile, which can be used for cross-field comparisons. Some bibliometricians recommend mean-based citation indicators (e.g., the characteristic scores and scales method; see Glänzel & Schubert, 1988) and others deem percentiles the best alternative among the indicators (Bornmann, Leydesdorff, & Mutz, 2013). Further, different variants of mean-based and percentile-based indicators are currently in use (Todeschini & Baccini, 2016). No gold standards exist which can be recommended for the general application of impact indicators in research evaluation.

If it should be the aim of bibliometrics to develop indicators which gain general acceptance (among bibliometricians and beyond), it is essential to criticize the available indicators sharply. This is especially necessary for field-normalized indicators, because these indicators have – as a rule – a more complicated design than simple indicators such as mean citation rates. We propose that indicators should be continuously investigated as to their reliability and validity against the backdrop of Popper's critical rationalism (Gingras, 2016;Lindner, Torralba, & Khan, 2018; Ulrich, 2006): bibliometricians should try to create or find situations in which the indicators possibly do not show the desired behavior. For example, an indicator which is intended to reflect the quality of publications should be correlated with the assessments of peers (Bornmann & Marx, 2015). If the correlation turns out to be low, the validity of the indicator can be questioned. Situations should be created or found in empirical research in which the measurement with the new indicator can fail. An indicator is regarded as preliminarily valid if these situations could not be found or realized. Such an indicator should gain general acceptance in citizen and professional bibliometrics.

In bibliometrics, we frequently find inadequate investigations of the validity of indicators in the literature. Good examples are the developments of h index variants since the introduction of the h index in 2005 (Hirsch, 2005). Since then, about 50 variants have been introduced (Bornmann, Mutz, Hug, & Daniel, 2011). In many of the corresponding papers, one or several weaknesses of the h index are unrevealed, before the new variant is introduced (see, e.g., Tol, 2009). Then, it is demonstrated on the base of one or several examples that the new variant performs (significantly) better than the original index. However, these tests do not guarantee that the indicators have been proved successfully. The authors seem to be more interested in producing favorable results for the new indicators than their empirical validation. For critical validation, situations should be sought in which the indicators probably fail and these situations should be specifically investigated. With the elimination of failed indicators, the set of indicators is reduced and standard indicators emerge.

For the critical examination of (field-normalized) indicators, some empirical and theoretical tests have been proposed which are explained in the following two sections. Since the success of field-normalization depends on the used field categorization scheme, we discuss the current (most frequently) used schemes in Section 4. Section 5 introduces the Werther effect describing the situation in which indicators become popular although theoretical and empirical results question their use.

2. Empirical tests of field-normalized indicators

For the examination of the convergent validity, the correlation between the field-normalized indicator and the assessment by peers is calculated (Bornmann & Marx, 2015). The backdrop of the test is that both methods of research evaluation measure the same construct, namely the quality of research (Thelwall, 2017). The peer review process is the oldest and most accepted method for research evaluation (Bornmann, 2011). If the correlation coefficient is substantial, the field-normalized indicator seems to measure research quality validly. The substantial coefficient should be at least at the medium level (about r = 0.5). We cannot expect a high correlation (about r = 0.7), since citations only measure one aspect of quality, namely impact. Importance, for instance, can scarcely be measured by citations (Martin & Irvine, 1983). The coefficient (r = 0.5) which should be reached by the correlation between field-normalized indicators and assessments by peers is a rough estimation, since the correlation can be performed on different levels of aggregation: correlations will usually be lower at the level of individual publications than at the level of research groups, departments, or institutes.

At least three further tests have been proposed which investigate the ability of indicators to minimize the effect of field-differences in comparison (e.g. mean-based citation indicators compared to percentile indicators). A valid field-normalized indicator adjusts the citation impact of a paper to the average field level and allows meaningful cross-field comparisons. The three tests require a field categorization scheme (see Section 4) which is independent of the scheme used for field normalization (Sirtes, 2012;Waltman & van Eck, 2013). This is also the disadvantage of these tests: one needs a second independent subject category scheme, which is able to reflect scientific fields properly. If such a scheme is not available, the tests cannot be applied.

The first test can be called "fairness test" (Bornmann, de Moya Anegón, & Mutz, 2013; Kaur, Radicchi, & Menczer, 2013; Radicchi, Fortunato, & Castellano, 2008). According to Radicchi and Castellano (2012) "the 'fairness' of a citation indicator is . . . directly quantifiable by looking at the ability of the indicator to suppress any potential citation bias related to the classification of papers in disciplines or topics of research" (p. 125). For example, Bornmann and Haunschild (2016a, 2016b) used the fairness test to study field-normalized indicators in the area of bibliometrics (for the testing of the indicator: citation score normalized by cited references, CSNCR) and altmetrics (Mendeley reader indicators). Other applications have been reported in the bibliometric literature.