FEMS Microbiology Letters, 365, 2018, fny059

doi: 10.1093/femsle/fny059
Advance Access Publication Date: 12 March 2018
Minireview

MINIREVIEW – Incubator

What are we measuring? Refocusing on some fundamentals in the age of desktop bibliometrics

Ian Rowlands†

Research Management & Innovation Directorate, King's College London, 150 Stamford Street, London SE1 9NH, UK

*Corresponding author: Research Management & Innovation Directorate, Franklin-Wilkins Building, King's College London, 150 Stamford Street, London SE1 9NH, UK. Tel: +44-7836-8656; E-mail: ian.rowlands@kcl.ac.uk

One sentence summary: Massive choice in new bibliometric indicators is creating new opportunities for evaluating research but it is also leading to a crisis of construct validity.

Editor: Dave Nicholas

†Ian Rowlands, http://orcid.org/0000-0002-0634-3376

ABSTRACT

The central challenge in bibliometrics is finding the best ways to represent complex constructs like 'quality,' 'impact' or 'excellence' using quantitative methods. The marketplace for bibliometric data and services has evolved rapidly and users now face quite unprecedented choice when it comes to the range of data now available: from traditional citation-based indicators to reader ratings and Wikipedia mentions. Choice and ease of access have democratised bibliometrics and this is a tool now available to everyone. The era of 'desktop bibliometrics' should be welcomed: it promises greater transparency and the opportunity for experimentation in a field that has frankly become a little jaded. The downside is that we are in danger of chasing numbers for numbers' sake, with little understanding of what they mean. There is a looming crisis in construct validity, fuelled by supply side choice and user-side impatience, and this has significant implications for all stakeholders in the research evaluation space.

Keywords: bibliometrics; research evaluation; construct validity

THE CONTEXT FOR BIBLIOMETRICS

According to the United Nations Educational, Scientific and Cultural Organisation (UNESCO) Science Report, gross global expenditure on research and development by government and non-government sources amounted to 1.48 trillion PPP (purchasing parity dollars) in 2013 (UNESCO 2015). This staggering sum of money supports the careers of nearly 7.8 million researchers and leads, by simple back-of-envelope calculations, to the publication of a peer-reviewed paper every 18 s.

These figures are invoked simply to put what follows into context.

For all the evident weaknesses of a metric approach to defining quality or excellence, the trouble is that peer review, the backbone of modern science, is far from perfect. Many studies have revealed poor levels of agreement between peer reviewers (for example, Allen *et al.* 2009; Eyre-Walker and Stoletzki 2013; Bertocchi *et al.* 2015; Bornmann 2015a). Eyre-Walker and Stolezki argue that as well as being 'error-prone, biased and expensive,' reviewers are particularly poor at judging a paper's likely future impact (as measured by subsequent citations). The volume of outputs is such that we have absolutely no choice but to apply scientific methods—including the sensible use of quantitative indicators—to policy formation, management and evaluation of this massive enterprise, but that is not to say that this is easy, and certainly not without its dangers:

> 'Since the first decade of the new millennium, the words *ranking*, *evaluation*, *metrics*, *h-index*, and *impact factors* have wreaked havoc in the world of higher education and research. Governments and

Received: 24 December 2017; **Accepted:** 8 March 2018

Table 1. Some advantages and disadvantages of bibliometrics for research evaluation.

Advantages	Disadvantages
Objective and replicable	Theoretically weak
Easily scaleable from a single researcher to a whole country	Highly platform-specific
Comprehensive coverage possible in some areas	Unsuitable in many subject areas
Broadly correlates with peer assessment	Author and institution names not well controlled
Sophisticated tools available	Backward-looking in time
Relatively inexpensive	Specialist expertise needed

> research administrators want to evaluate *everything*—teachers, professors, researchers, training programs, and universities—using quantitative indicators (Gingras 2016:vii).'

Yves Gingras has written a considered polemic about the uses and abuses of bibliometrics in research evaluation, from which quote above is taken, and he asks some very serious questions both about the validity of many of the indicators that are in common circulation and the unreflexive way they are often used.

Issues of meaning assume even greater relief now that we are being exposed to a massive choice of new metrics: from mentions in policy documents to citations to data sets, from reader ratings to new measures of journal impact based on network theory, all available at the click of a mouse. What used to be a highly arcane and specialist endeavour requiring hours of manual effort and a pointed hat is now becoming a mainstream activity for many in the research business, with the arrival of 'desktop bibliometrics.'

Like all research methods, bibliometric techniques have considerable advantages and disadvantages for research assessment (see Table 1):

The key advantages are that we can bring open, transparent and replicable methods to bear on questions relating to the effectiveness of research and innovation systems, regardless of scale, and we can do so quickly and efficiently. Moreover, we have amassed a considerable knowledge base, with around 1500 peer-reviewed papers in good journals published every year on bibliometrics and scientometrics, and we have deep reserves of expert human capital to draw on. Still, for the present writer, a 'reluctant bibliometrician,' there is a feeling that we are falling badly behind the curve in some respects and that we need to confront some rather difficult questions full on—the kinds of questions that people inside the research assessment enclosure would probably prefer not to hear. I will argue that the vast expansion of different metrics is only serving to highlight existing problems in construct validity that are now beginning to assume serious proportions. All this is highly relevant to the recent buzz around 'responsible metrics' and this review concludes with some recommendations to different parts of the academic community in terms of how we adapt and flourish in the age of desktop bibliometrics.

AN EXPLOSION OF CHOICE

The origins of bibliometrics lie in serials collections management, an early example of evidence-based professional practice, where librarians used quantitative methods for example to help decide which journals to subscribe to, and which to discontinue (Bradford 1934). During the 1980s, interest shifted to questions of evaluating researcher productivity and impact, but most usually at high levels of aggregation: such as helping to understand the development and relative performance of national science systems, or to benchmarking leading institutions. Much of the data collection involved access to expensive proprietary data sets and considerable manual effort to clean up and present the findings.

Table 2. The range and diversity of publication metrics available today.

Scholarly activity	Information sources
Software downloads	Codeplex, Bitbucket, Launchpad
Repository full text downloads	Individual institutional and subject repositories, and aggregator services such as IRUS-UK
Article full text downloads	Publishers such as Public Library of Science (PLoS), Springer
Wikipedia mentions	Wikipedia, PlumX, Altmetric.com
Monograph sales and ranking	Amazon, Nielsen BookScan, Sales Rank Express, NovelRank
Monograph holdings	WorldCat, Copac, PlumX
Citations to books and book chapters	Google Scholar, InCites, Microsoft Academic, SciVal
Journal acceptance rates	Cabell's Directory of Publishing Opportunities, individual journal web sites
Reader ratings and reviews	Amazon, Goodreads, PlumX
Journal impact factors (broadly defined)	Clarivate, CWTS Journal Indicators, Elsevier SciVal, Scimago Journal & Country Rank
Data citations	Data Citation Index, Google Scholar
Article-level citation metrics	Clarivate InCites, Elsevier SciVal, National Institutes of Health iCites
Software citations	Depsy, Google Scholar
News mentions	Altmetric, Newsflow, PlumX
Mendeley readers	ImpactStory, Scopus
Post-publication peer review and recommendations	F1000, Pubpeer
Policy mentions	Altmetric

The past five years have seen an incredible supply side expansion in the breadth and diversity of publication metrics now available for analysis. As well as the more traditional citation-based metrics, we now find services that offer almost any conceivable quantification of scholarly activity, dissemination and impact. Table 2 enumerates some of the research metric information services that are currently available at the time of writing this mini review.

Many of these services are free, or at least directly available to members of the academic community through centrally funded subscriptions. As a result, bibliometrics has become democratised and this has enabled researchers, managers, administrators and funders to conduct their own desktop bibliometrics, often digging down to an individual, or even a subset of an individual's output.

As well as a proliferation of data services, there has been a massive expansion of the bibliometric vocabulary. A major driver, but by no means the only one, for this proliferation of metrics is the availability of alternative ('alt') or complementary

metrics based on interactions with social media. Academics are increasingly finding serious application for social media across all points of the communication lifecycle (Rowlands *et al.* 2011) and understanding scholarly communication without reference to this context is scarcely imaginable any longer. Services like Altmetric and Plum X offer exciting potential and it is hardly surprising that they have become such hot topics in scientometrics, especially as they shed light on non-formal communication, such as between academics and practitioners, and so 'the future, then, could see altmetrics and traditional bibliometrics presented together as complementary tools presenting a nuanced, multidimensional view of multiple research impacts' (Priem, Piwowar and Hemminger 2012:1). As might be expected, social media and citation counts vary in the degree to which they are correlated (Bornmann 2015b) raising questions about which communication dynamics they are measuring, and even more interesting questions about what this means for our understanding of 'research impact.'

A recently published handbook of bibliometric indicators (Todeschini and Baccini 2016), essentially a technical recipe book, runs to more than 500 pages. In it you will find 24 different recipes for calculating the fractional contribution of researchers to a multi-authored paper, more than 40 variants of the *h*-index, together with indicators you would never have dreamt of. Take the Knudop effect, for instance. This attempts to quantify the additional citation credit that flows unfairly to authors from high-prestige institutions (or the Podunk effect, its opposite, which measures the citation handicap associated with authorship from a low-ranking institution). Or take the *q*-index (Bartneck and Kokkelmans 2010). This measures the propensity of an author to strategically self-cite their earlier papers in order to nudge up their *h*-index: a classic example of Goodhart's Law in action.

All quantitative indicators are brutally simplistic representations of reality, with evident failings, and bibliometricians are much given to proposing incremental technical 'improvements,' fiddling around the edges. The product of all these forces is that never before have we had so much access, so easily, to so many research indicators or the tools to access and process them. In turn, this is confusing rather than illuminating matters. Do you want a journal impact factor? Well you now have considerable choice. In addition to the 'classic' journal impact factor (which comes in two- or five-year versions), Elsevier recently coined CiteScore, a journal-level metric that closely mimics the classic factor with a few tweaks but much wider coverage. Or you can choose SNIP (Source Normalised Impact per Paper) that corrects for field differences in citation rates. Or Scimago Journal Rank, a highly sophisticated metric that takes account both of the number of citations received by a journal and the prestige of the journals from where the citations originate. The question I dread, as many of my colleagues must, is 'Which is the best?' This takes us into the really difficult territory of what these indicators actually measure?

WHAT ARE WE MEASURING?

Bibliometric indicators such as citation or social media attention metrics are often held out as a proxy for the quality or impact of a research output. As we have seen, we have an enormous, and confusing, array of independent variables to choose from, but what is their relation to the relevant dependent variable: quality or impact? These are clearly complex and multi-dimensional concepts.

In terms of trying to pin down what a high-quality piece of research might look like, we might well consider something like the output quality criteria developed for the purposes of the 2014 UK Research Excellence Framework:

(i) Scientific rigour with regard to design, method execution, and analysis;
(ii) Addition to knowledge and to the conceptual framework of a field;
(iii) Potential and actual significance of the research;
(iv) Logical coherence of the argument;
(v) Contribution to theory building;
(vi) Scale, challenge, and logistical difficulty posed by the research;
(vii) Significance of work to advance knowledge, skills, understanding and scholarship in theory, practice, management and/or policy;
(viii) Applicability and significance to the relevant service users and research users;
(ix) Potential applicability for policy in, for example, health, healthcare, public health, animal health, or welfare. HEFCE (2012:42–43)

In the particular context of a national assessment process, 'high quality' is clearly taken to mean rigour, originality and significance, with additional credit for research that furthers wider social, economic or cultural goals.

Finding ways to represent the qualities in the list above numerically using proxy measures is the central challenge in bibliometrics, as well as its Achilles' heel for its many detractors. Citation (or more accurately reference) counts are the bedrock of evaluative bibliometrics, so we first ought to consider where and how those counts arise. David Shotton (2010) has developed a machine-readable ontology (CiTO) for the semantic web that teases out the fundamental nature of reference citations in the scientific literature. He finds 23 factual and rhetorical relationships between citing and cited documents (see Table 3).

It should be completely obvious from Shotton's analysis that we are on shaky intellectual ground indeed if we simply aggregate counts of citations and then project them as an accurate representation of quality or impact. This does little justice to the complex knowledge networks that exist within the scientific literature, linking papers, authors and research projects. The CiTO ontology is published under a Creative Commons attribution licence (Shotton 2010) and could potentially form the basis for a much more nuanced and sophisticated understanding of what reference citations actually mean—in context. Of course, the practical challenges of applying semantic web techniques on any scale to the scientific literature are profound, not least the fact that 'citations are not usually freely available to access, they are often subject to inconsistent, hard-to-parse licenses and they are usually not machine-readable' (Initiative for Open Citations, 2017).

A landmark study by Steven Greenberg shows the value of citation analysis at this fine level of detail and sophistication (Greenberg 2009). By characterising 675 citations in 202 papers on a hypothesised relationship between inclusion body myosotis and the accumulation of ß-amyloids in the brain in Alzheimer's disease, Greenberg came to come some very uncomfortable conclusions for citation bean counters. He found evidence of unfounded authority—papers that supported the hypothesis were much more likely to be cited than those that refuted it. Many of the cited papers amplified the rhetorical force behind the hypothesis, while actually not providing any new evidence. Of most concern was his finding that early papers setting

Table 3. The 23 relationships between citing and cited document in CiTO (from Shotton 2010).

	Rhetorical relationships		
	Positive	Negative	Neutral
cito:cites	cito:confirms	cito:corrects	cito:discusses
cito:citesAsAuthority	cito:credits	cito:critiques	cito:reviews
cito:citesAsMetadataDocument	cito:extends	cito:disagreesWith	
cito:citesAsSourceDocument	cito:obtainsSupportFrom	cito:qualifies	
cito:obtainsSupportFrom	cito:supports	cito:refutes	
cito:supports	cito:updates	cito:corrects	
cito:usesDataFrom		cito:critiques	
cito:usesMethodIn		cito:disagreesWith	
cito:cites			

out the hypothesis became morphed into statements of 'fact' in later citing papers. So, simple counts of references may not be such a good guide to quality after all and perhaps citations are not quite as objective as many would have us believe.

HOW DO WE KNOW IF OUR MEASUREMENTS ARE ANY GOOD?

One of the key problems in bibliometrics lies in its theoretical weakness. More specifically, there is a current crisis in construct validity that will eventually bring the whole edifice down if we it is not tackled soon. Construct validity refers to the degree to which a test measure (e.g. a reference count or an Altmetric score) measures what it claims or purports to be measuring (e.g. quality or social engagement)?

One, admittedly fairly indirect, test of the construct validity of bibliometric indicators is how strongly they associate statistically with peer review judgments? A number of studies in the UK have looked for and found evidence of a broad correlation between the outcome of national assessment exercises (the 2008 Research Assessment Exercise and the 2014 Research Excellence Framework) and citation metrics (inter alia; Norris and Oppenheim 2003; Myrglod *et al.* 2015; Wilsdon *et al.* 2015; Wooding *et al.* 2015). At an aggregate level, there does seem to be considerable evidence for medium to high levels of correlation between peer assessment and citation measures in some subject areas (notably Clinical Medicine), but the problem is that rankings based solely on peer review or metrics will result in significantly different outcomes. A study comparing quantitative (citations, journal impact factors) and qualitative measures (expert review and F1000 reader ratings) for 979 Wellcome Trust funded biomedical papers came to a similar conclusion: that 'bibliometric measures may not be sufficient in isolation as measures of research quality and importance, and especially not for assessing single papers or small groups of research publications' (Allen *et al.* 2009). A strong consensus in the responsible bibliometrics community now favours a 'variable geometry' based on expert judgment, quantitative and qualitative indicators, recognising that all methods, even peer review, have their flaws.

The trouble is, desktop bibliometrics are cheap, easy and give instant results, and being realistic, people simply do not have the time to read large numbers of papers, or even to organise their reading by others on sufficient scale. The danger is that simplistic measures are deployed without the user always understanding their real meaning and validity. Worse still, we may be seeing Goodhart's Law in action: 'when a measure becomes a target, it ceases to be a good measure.' What begins as a well-intentioned exercise in using research indicators to help assess quality can easily morph into a situation where maximising citation counts or social media attention, becomes the convenient end goal in itself, rather than fundamentally addressing issues of research quality or sustainability.

Samuel Messick (1995) has developed a comprehensive unified framework for construct validity in the context of psychological testing, that is highly pertinent to questions of what we are measuring, and how, in relation to research assessment. Messick argues that there are six questions we should ask to confirm the construct validity of a test, or in this case the application of research metrics in a given situation. Basically, to what extent do the metrics actually measure what they purport to measure?

Question 1: What are the consequential risks if the indicators used turn out to be invalid or inappropriately interpreted? Is the test worthwhile, given the risks? This could well be a serious issue in relation to the inept use of bibliometrics in recruitment or promotion processes.

Question 2: Do the indicators properly measure the content of the dependent variable we are interested in? In the case of citation or other publication metrics, the fit with quality or impact must at best be partial.

Question 3: Is there a substantive theoretical foundation underlying the specific choice of indicator? Again, unlikely in most cases, bibliometrics is an opportunistic activity.

Question 4: Does the bibliometric indicator exhibit structural validity? This concept is often used in the context of questionnaire design: To what extent does the measurement scale used adequately reflect the dimensionality of the construct? A search for a mention of this concept in the bibliometrics literature has proved fruitless: perhaps the reason is simply that we are generally so unclear what we are measuring that this question goes unasked and unanswered?

Question 5: To what extent does a bibliometric exercise exhibit externally convergent and discriminant qualities? In other words, does the indicator satisfy the condition that it is positively associated with the construct that it is supposed to be measuring (i.e. convergent)? The criteria for convergent validity would not be satisfied in a bibliometric experiment that found little or no correlation between, say, peer review grades and citation measures. This is one of the reasons why bibliometrics are simply inappropriate in some fields, like political science, even if one discounts the limited scope and coverage of many databases. If an indicator has discriminant validity, then it should most certainly not associate with two completely unrelated or opposite constructs. In a blistering demolition of the

construct validity of the *h*-index Barnes (2016) points out that an indicator that purports to measure impact in effect functions as a measure of output—the number of papers, with a very high correlation. So, to return to the theme, are we clear precisely what are we measuring?

Question 6: How far do the results of a bibliometric exercise generalise across different groups, settings and tasks? A very difficult question this because we know that publication and related scholarly practices reveal enormous age and field differences. More positively, there is a great deal of comparative evidence, and theorising (e.g. Hjørland 2002) on these 'domain differences' in the literature. This territory may be familiar to information scientists and bibliometricians, but subject differences can lead to serious iniquities if the indicators are presented without field sensitivity.

Aside from these philosophical matters, there is much empirical evidence that 'para-textual' factors unrelated directly to article quality (such as number of authors, page length and the number and citation impact of references) subtly influence citation rates (Bornmann and Leydesdorff 2015). Few studies make allowance for these kinds of confounding factors.

Questions of validity and confounding factors are important, and they urgently need to be revisited given the plethora of new metrics and services on offer. Harnad (2008) makes the point very well: '[Metrics] need to be jointly tested and validated against what it is that they purport to measure and predict, with each metric weighted according to its contribution to their joint predictive power. The natural criterion against which to validate metrics is expert evaluation by peers' (Harnad 2008:103). The latter is admittedly an expensive commodity, and one that is very difficult to scale, but other secondary analysis methods may offer a way forward: for instance using content analysis to better understand the context within which a research output is being tweeted (Wilkinson and Thelwall 2012).

IMPLICATIONS FOR STAKEHOLDERS IN BIBLIOMETRICS

Research metrics are becoming increasingly dominant in our universities and funding agencies, so it behoves us to use them responsibly and wisely. In this final section, I consider some positive measures that various stakeholders in the bibliometric environment should adopt take to make the best of a golden opportunity.

Journal publishers could help support to greater transparency by lending their support to the OpenCitations project, a movement that seeks to tear down the citations paywall and make scholarly citation links freely available under licensing conditions that would facilitate the creation of non-proprietary citation services. Many publishers are releasing article-level data on their full text downloads, but it would be great to see more doing this, and even the creation of aggregation services, so that we can see the bigger picture.

Suppliers of bibliometric services need to more open about the limitations of their data and be fully transparent—in detail—about how their indicators are calculated, and how they should be interpreted: ideally in terms of explicit confidence intervals appropriate for the sample size.

Institutions need to consider how and when (and when not) to use quantitative indicators in their internal management and evaluation processes. Some universities have started to develop their own policies on the responsible use of metrics, following one of the key recommendations of The Metric Tide report (Wilsdon *et al.* 2015). In addition, many are signing up to DORA (2013), the San Francisco Declaration on Research Assessment in order to reinforce the point that judging an article by its cover (the journal impact factor) is simply wrong-headed and unnecessary when we have easy access to an array of article-level metrics.

Managers, administrators and researchers need to spend a little time to become better informed about bibliometrics and as informed consumers so that they can make better and more nuanced decisions. Crucially, they need to be reminded that metrics have little face validity and can only partially capture notions of quality or excellence. Academic folklore is full of persistent myths about citations. Most papers are never cited—wrong! Most papers do get cited eventually, it just depends how long a view you take of the market—in fact the proportion of uncited papers has fallen dramatically since the 1970s (Larivière, Archambault and Gingras 2008). Papers are not cited after about 5 years—wrong! In fact, the median age of references in scientific papers has increased significantly over the past three decades (Wallace, Larivière and Gingras 2008). Most importantly, the ultimate users of research metrics and analysis need to be more realistic about drawing major conclusions from small samples of papers (see Williams and Bornmann 2016, for an interesting discussion on sampling issues in bibliometrics and the inadequate power of many analyses). They should also consider alternatives to what Joseph Schneider calls the 'null ritual': the unthinking blanket application of null hypothesis significance tests (Schneider 2015). Slavery to arbitrary thresholds of p to test for statistical significance provides far less useful information than effect sizes and confidence intervals.

Professional services staff who deal with research metrics on a day-to-day basis need to be very careful how they present their analyses, and place appropriate limits on their work. In particular, they need to consider sample size and power, and place confidence intervals around the mean when dealing with ratio data (Rowlands 2017). They clearly also have a significant role to play in raising levels of metrics literacy within their organisations

Finally, academic researchers with an interest in scholarly communication could serve the whole community well by helping us to better understand the construct validity of bibliometric indicators and which aspects of research quality or impact are best captured. We tend to think of quality in terms such as originality, significance and rigour. Is it the case that citations, for instance, are better at measuring significance than rigour? Or perhaps social media indicators are better at spotting originality? Who knows, but these are interesting and rather fundamental questions to be asking at this point in time.

Conflict of interest. None declared.

REFERENCES

Allen L, Jones C, Dolby K *et al.* Looking for landmarks: the role of expert review and bibliometric analysis in evaluating scientific publication outputs. *PLoS One* 2009;**4**:e5910.

Barnes CS. The construct validity of the *h*-index. *J Doc* 2016;**72**:878–95.

Bartneck C, Kokkelmans S. Detecting *h*-index manipulation through self-citation analysis. *Scientometrics* 2011;**87**:85–98.

Bertocchi G, Gambardella A, Jappelli T *et al.* Bibliometric evaluation vs. informed peer review: evidence from Italy. *Res Policy* 2015;**4**:451–66.

Bornmann L. Interrater reliability and convergent validity of F1000 Prime peer review. *J Assoc Inf Sci Tech* 2015a;**66**:2415–26.

Bornmann L. Alternative metrics in scientometrics: a meta-analysis of research into three altmetrics. *Scientometrics* 2015b;**103**:1123–44.

Bornmann L, Leydesdorff L. Does quality and content matter for citedness? A comparison with para-textual factors and over time. *J Informetr* 2015;**9**:419–29.

Bradford SC. Sources of information on specific subjects. *J Inf Sci* 1934;**137**:85–6.

Eyre-Walker A, Stoletzki N. The assessment of science: the relative merits of post-publication review, the impact factor, and the number of citations. *PLoS Biol* 2013;**11**:e1001675.

Gingras Y. *Bibliometrics and Research Evaluation: Uses and Abuses*. Cambridge MA: MIT Press, 2016.

Greenberg SA. How citation distortions create unfounded authority: analysis of a citation network. *BMJ* 2009;**339**: b2680.

Harnad S. Validating research performance metrics against peer rankings. *Ethics Sci Environ Polit* 2008;**8**:103–7.

HEFCE. *Research Excellence Framework 2014: Panel Criteria and Working Methods (REF 01.2012)*. Bristol: Higher Education Funding Council for England, 2012.

Hjørland B. Domain analysis in information science. *J Doc* 2002;**58**:422–62.

Initiative for Open Citations (I4OC). 2017. Available at: http://io4c.org (26 March 2018, date last accessed).

Larivière V, Archambault E, Gingras G. Long-term variations in the aging of scientific literature: From exponential growth to steady-state science (1900–2004). *J Am Soc Inf Sci Tec* 2008;**59**:288–96.

Messick S. Validity of psychological assessment: validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *Am Psychol* 1995;**50**:741–9.

Myrglod O, Kenna R, Holvatch Y *et al*. Predicting results of the Research Excellence Framework using departmental *h*-index. *Scientometrics* 2015;**102**:2165–80.

Norris M, Oppenheim C. Citation counts and the Research Assessment Exercise V. *J Doc* 2003;**59**:709–30.

Priem J, Piwowar HA, Hemminger BM. Altmetrics in the wild: using social media to explore scholarly impact. 2012. Available at: http:arXiv:1203.4751v1 (26 March 2018, date last accessed).

Rowlands I. SciVal's field-weighted citation impact: sample size matters! *Bibliomagician: comment and practical guidance from the LIS-Bibliometrics community*. [Blog post]. 2017. Available at: https://thebibliomagician.wordpress.com/2017/05/11/scivals-field-weighted-citation-impact-sample-size-matters-2/ (26 March 2018, date last accessed).

Rowlands I, Nicholas D, Russell B *et al*. Social media use in the research workflow. *Learn Publ* 2011;**24**:183–95.

DORA. *San Francisco Declaration on Research Assessment*. 2013. Available at: http://sfdora.org/read (26 March 2018, date last accessed).

Schneider JW. Null hypothesis significance tests. A mix-up of two different theories: the basis for widespread confusion and numerous misinterpretations. *Scientometrics* 2015;**102**:411–32.

Shotton D. CiTO, the citation typing ontology. *J Biomed Sem* 2010;**1**:S6.

Todeschini R, Baccini A. *Handbook of Bibliometric Indicators: Quantitative Tools for Studying and Evaluating Research*. Weinheim: Wiley-VCH Verlag, 2016.

UNESCO. Science Report: Towards 2030. Paris: UNESCO, 2015, ISBN: 978-3-527-33704-0.

Wallace ML, Larivière V, Gingras Y. Modeling a century of citation distributions. *J Informetr* 2009;**3**:296–303.

Wilsdon J, Allen L, Belfiore E *et al*. *The Metric Tide: Report on the Independent Review of the Role of Metrics in Research Assessment and Management*. Bristol: Higher Education Funding Council for England, 2015.

Wilkinson D, Thelwall M. Trending Twitter topics in English: an international comparison. *J Am Soc Inf Sci Tec* 2012;**63**:1631–46.

Williams R, Bornmann L. Sampling issues in bibliometric analysis. *J Informetr* 2016;**10**:1225–32.

Wooding S, Van Leeuwen TN, Parks S *et al*. UK doubles its "world-leading" research in life sciences and medicine in six years: Testing the claim? *PLoS One* 2015;**10**: e0132990.