
Book Section:
Booth, A. orcid.org/0000-0003-4808-3880 (2016) "Metrics of the trade": where have we come from? In: Tattersall, A., (ed.) Altmetrics: A practical guide for librarians, researchers and academics. Facet Publishing , London , pp. 21-48. ISBN 9781783300105

“Metrics of the trade”: where have we come from?

Introduction

Paradoxically with altmetrics in the ascendancy the scholarly world is witnessing a renaissance in interest in traditional metrics. This chapter examines longstanding and recent drivers for metrics from the complementary perspectives of scientists, research organisations and funding agencies. It outlines established metrics for individuals (e.g. citations and the h index) and for teams and journals (the journal impact factor and other proprietorial alternatives). It examines the extent to which these metrics correlate to other important characteristics such as newsworthiness and journal prestige.

The chapter considers challenges raised by the emerging “impact agenda” including a need to capture “impact on society, social impact, real-world impact, knowledge translation, and uptake by the public” (Eysenbach, 2011). It also describes how both traditional and impact metrics have been used within a variety of research and performance management contexts, giving examples of appropriate and inappropriate usage.

The chapter examines criticisms of established measures and how these criticisms might, at least in theory, be addressed. It discusses manipulation and game playing, together with high profile examples, which exploit acknowledged weaknesses of metrics. It examines the extent to which relatively recent forms of publishing, such as open access journals, are accommodated by, or pose challenges to, traditional metrics.

The chapter sets the scene for the considered evaluation of altmetrics that follows in subsequent chapters. It concludes by examining whether traditional metrics still occupy a role within a world increasingly populated by social media and social networks.

On Metrics and Madness

In Prokofiev’s Lieutenant Kijé, a slip of a clerk’s pen on a list of officers for promotion compiled for a mad Tsar, leads to the creation of the fictitious hero. Fearful of displeasing the mad Tsar his advisers manufacture increasingly elaborate escapades for the imaginary Kijé, each rewarded by a successive promotion. Finally, following Kijé’s promotion to General the Tsar wants to meet the hero whose career he has followed with interest. His alarmed courtiers “kill off” the Tsar’s protégé and manufacture the death and funeral of the now-General Kijé. In 1986, in an instance of life imitating art, De Lacey and colleagues (1985) revived an account from the 1930s (Dobell, 1938) of a classic error of this type. The title of a paper published in a Czechoslovakian medical journal in 1887 began with "O uplavici", the Czech (Bohemian) phrase for “On Dysentery”. In error this phrase was transcribed as the author's name, Uplavici O, by an abstracter. Like Kije the fictional O Uplavici enjoyed prodigious longevity surviving for some 50 years and even acquiring a doctorate from an American indexer in 1910!

Like the mad Tsar from Prokofiev's opus, the academic community continues to attach unwarranted significance to the metrics of its printed outputs. Many promotions are attributable to the interpretation or, indeed misinterpretation, of such metrics and, no doubt this is equally true of the occasional doctorate. A Carnegie Foundation study (Boyer, 1990) reported that universities in the U.S. typically count citations or publications when reviewing their faculty for tenure, promotions, demotions, merit increases, etcetera. While the time for "killing off" traditional metrics lies in the unspecified future this chapter seeks to document the past and present of established academic metrics[1]. In doing so we hope to inform ongoing critical use of an emerging battery of altmetrics.

Drivers for metrics

For centuries peer review has been the vehicle for determining qualitatively whether research is appropriate, represents good value for money, and can make a useful contribution to society (Reedijk, 1998). However as van Raan (2003) observes: "Opinions of experts may be influenced by subjective elements, narrow-mindedness and limited cognitive horizons". Bibliometrics is seen as one way of addressing such limitations. However, as van Raan demonstrates, this may simply exchange one set of imperfect judgements for another. The creation of the Science Citation Index (SCI) in 1961 was the catalyst for the emergence of citation analysis as an independent field of study (MacRoberts & MacRoberts, 1989). Initially the Institute of Scientific Information (ISI) was very proprietorial about its innovative ranking systems. In recent years alternatives have included the Scopus system, preferred by the UK Research Excellence Framework, and Google Scholar which, because of its broader "more democratic" reach, typically demonstrates higher citation counts and more incorrect citations (Harzing & van der Wal, 2008).

Numerous drivers underpin the relentless move towards scientometrics, including bibliometrics. From a political perspective there is increased pressure for judicious use of public monies as universities and research centres seek to demonstrate scientific performance and wider societal impact. Use of seemingly objective measures gives the public a more transparent picture of scholarly activity. Economically the knowledge economy requires research to offer return on investment. Metrics also fuel competition amongst academics, allegedly with a positive effect on scientific quality and productivity. Socially, universities and other academic institutions seek to provide evidence that they are making an "impact", particularly in bringing to bear research findings upon pervasive societal problems. Technologically, the publication chain is easier to monitor and to measure, from early sharing of protocols and proposals through to submission of the archaically titled "manuscript", its subsequent dissemination and resultant social network activity. Underpinning the above is a belief in the collective wisdom of crowds with the implication that the judgements of an academic community are less subjective than that of the individual researcher themselves.

Amateur Bibliometrics

Van Raan (2005) appropriately characterises the use of citation and publication counts in determining academic progression as "amateur bibliometrics". This counting culture is predicated on the belief that bibliometric techniques can somehow

[1] Unless specifically mentioned 'metrics' is used to include both bibliometrics and, more broadly, scientometrics, although most examples used in this chapter derive from the former.

measure the “otherwise elusive concepts of quality and influence” (Schoonbaert et al, 1996). Objections to this assumption lie in how such metrics are calculated and, more significantly, in how they are misinterpreted and misapplied. Three brief and diverse examples illustrate such limitations. The Impact Factor, an established and manufactured metric, relates the number of citations to the number of citeable articles within a two year period. Why two years? Many studies take more than two years to make their academic mark. Indeed in the pre-Web era it would take more than two years to discover that a paper existed, let alone to publish a paper that cites it. Is a journal’s reputation to be assessed within such a transitory evaluation period?

The h index – An author has index h if h of his or her total number of papers have at least h citations each and the remaining papers have ≤h citations each. Therefore an h index of 26 indicates that 26 papers by that author have been cited 26 or more times (Hirsch, 2005).

As another example the h index, a metric that seeks to capture both quality and quantity of academic outputs, is calculated from any authorial contribution regardless of the author’s position in the author order. A lead author with a handful of publications, each cited hundreds of times, could acquire an h index inferior to that for a research assistant who has contributed to twenty or more modestly cited publications as 7th author. Finally numbers of citations vary significantly by discipline, regardless of how problematically “discipline” is defined. The relative performance of the same article differs according to whether its authors decide to publish it in a journal of its own discipline, in another discipline or in a generalist journal, such as the British Medical Journal within medicine. Furthermore the number of citations is affected by such extrinsic considerations as editorial policies regarding open access, the publishing house with which a journal is affiliated and even whether the chosen journal is included within a bundle of journals offered as a discount package to university libraries. Before we judge such flawed metrics too harshly we must recall that, prior to development of recent methods of research assessment, research funds were allocated according to raw numbers of students. We can at least celebrate the fact that the current regime no longer perpetuates “big is beautiful” in such an overt way although undoubtedly this mantra still exercises a covert influence.

The above examples reflect that traditional metrics operate at one of three levels:

1. In assessing individual performance
2. In evaluating the performance of a research team, group, department or institution
3. In recognising the contribution of a journal

A further dimension, to mention in passing, is whether metrics are designed to measure the performance of the researcher or of the research. However, taking the above levels in turn:

Assessing individual performance

Although individual academic performance is frequently measured through such measures as the number of articles published, the number of PhD students supervised, the value of research awards particularly as Principal Investigator, not to mention one's h index or the number of citations to individual articles the strength of these metrics derives from their multiplicity not their specificity. The value attributed to each individual metric may vary according to the priorities of a particular institution or even the time period within which particular targets are set. Metrics such as the h index are not sensitive to the current time period – they are uni-directional and only increase, never decrease. Within health services research an h index increases by one point for every year that an individual has operated as a full-time researcher. The h index is frequently disassociated with recent achievement or performance.

Other metrics carry intrinsic assumptions and thus favour particular models of research – for example the role of the Principal Investigator is clearly different in a culture of small specialist research teams when compared to large multidisciplinary research teams, where each co-investigator makes an idiosyncratic and valued contribution. Different disciplines observe different conventions in relation to ordering of author names within an article with differing interpretations placed upon the last named author. Authorial positioning may variously reflect, advancing alphabetical order, diminishing contribution or a political statement of research leadership – sometimes conflating the latter two factors! Medicine is witnessing is a move to acknowledging the extent of the contribution, to specify the exact nature of that contribution and to attribute most value to the first named author and to the corresponding author. Clearly traditional metrics differentiate poorly between different assumptions within differing research cultures, making the use of common metrics within the same university or their transferability across institutions persistently problematic.

Evaluating the performance of a research team, group, department or institution

The use of bibliometric measures to evaluate the performance of a research team or institution typically represents an imperfect and inadequate attempt to turn a very general indicator of performance into a specific performance indicator (PI) (Bence & Oppenheim 2005). In fact commentators make a useful distinction whereby PIs, unlike a simple indicator (such as an objective numerical figure), 'imply a point of reference… or comparator, and are therefore relative rather than absolute' (Bence & Oppenheim 2005). Experience from traditional metrics suggests that we need a more nuanced understanding of what is indicated by a particular performance measure if it is to be used in other than a crude and misleading way. To illustrate with the performance of two fictional research groups, Team Alpha and Team Numeric; Team Alpha employs a traditional "best appropriate choice" to selection of journals for their published outputs. If rejected by their first choice journal they go for the next appropriate journal and so on. In contrast Team Numeric selects initial target journals on the basis of their Impact Factor, and all things being equal, other efficiency considerations such as journal rejection rate and average time to publication. At the end of year five the performance of Team Numeric, unsurprisingly, outstrips that of its Team Alpha rival when evaluated simply in terms of citations and impact factors. Does this mean that Team Numeric is the better quality research team? In actuality we face three responses to this questionable

verdict: 1) Taking the indicator at face value we reward Team Numeric for their game-playing; 2) we discredit any assessments based on such a manipulable indicator, dismissing the metrics as “immature”; or 3) we mitigate the impact of this single measure, surrounding it with other metrics with similar inadequacies. Use of a battery of measures does not necessarily make a valid verdict more likely. Multiple measures may simply make it more problematic and time consuming to identify a single game-playing strategy!

Recognising the contribution of a journal

When authors choose the journal in which to publish they typically effect a compromise between aspiration and expectation. Generally an author seeks to publish in a journal with as high an impact factor as possible. However additional considerations include the journal’s particular niche, how the journal is viewed by a “college” of similar authors, the likelihood of rejection and the amount of time the author expends on “salvage” strategies, if rejected. A journal’s rejection rate may superficially appear as an informative measure of journal quality – in re-interpreting Groucho Marx’s quip about not wanting to “belong to any club that will accept people like me as a member” – yet it is clearly an independent measure of supply, both within a particular discipline and to a particular editorial office. Furthermore, there is a profound difference between acceptance for a print journal, which is partly determined by predetermined page budgets and subscription rates, and for an open access electronic journal, where editorial office capacity emerges as critical, given the comparatively negligible cost of additional electronic pages, supplements or online appendices. Qualitative assessments of a journal’s quality may relate to its longevity, the calibre of its editorial board, the reputation of its publisher, affiliation to a scientific organisation and the list goes on. More controversially judgements may depend upon where it hosts its editorial office and the pedigree of other journals in the publisher’s stable. Clearly it is problematic to determine an unambiguous cause and effect between such qualitative factors and the perceived “academic quality” of the journal title. It is even more challenging to relate a single metric, such as the Journal Impact Factor, as unequivocally reflecting quality. Correspondence in Nature (Dimitrov et al, 2010) describes how the journal Acta Crystallographica Section A experienced a meteoric increase in Impact Factor from 2.051 in 2008 to 49.926 in 2009 to leapfrog Nature (31.434) and Science (28.103) (Grant, 2010). Close analysis revealed that the article "A short history of SHELX", included the sentence: "This paper could serve as a general literature citation when one or more of the open-source SHELX programs…are employed in the course of a crystal-structure determination". As a consequence the article received more than 6,600 citations. This article became an incredible outlier in a journal where each article is cited on average three times, the second-most cited article in the same journal in the same year only had 28 citations. Of course the scale of this prodigious citation rate will attract further citations! Has the journal become 25 times better by accepting a single manuscript?

Table 1 - Illustrative Metrics by Domain

Category of Performance	Illustrative “metric”
Popularity	Number of Citations
Newsworthiness	Number of mentions in the media; column inches
Reputation	Number of Research Grants

	Value of research
Profile	Number of registered PhD students
Research supervision and management	Number/Percentage of completing PhD students (within specified time period)
	Average time to completion for PhD students
Accessibility and Popularity	Journal Impact Factor
	Normalised Citation Rates
Quantity/Quality	h index
Impact (Reach and Significance)	Impact Case Study

Relationship to quality

We acknowledge that, overall and on average, there is a relationship between traditional metrics and the likely quality of published outputs. Highly cited articles do tend to appear in high impact journals which do tend to attract the best research from the most distinguished authors. A large analysis of citations to articles in emergency medicine revealed that the citation count of articles was partly predicted by the impact factor of the journal in which they appeared and, to a more limited extent, by quality of the articles (Callaham et al. 2002). However there are important reservations to this statement. First, traditional metrics lack sensitivity to additional factors relating to academic quality. Conversely they may attach inappropriate value to factors at best irrelevant or, more alarmingly, subject to systematic error or bias. Thus a study examining citations to papers reporting randomized trials in hepato-biliary disease found a significant association with a positive outcome. However the study found no association of numbers of citations with adjudged quality (Kjærgard & Gluud 2002). Broadening the picture four different studies on levels of evidence in medical and/or psychological research emphasise the apparent inconsistency of results. Two studies of surgery journals found a correlation between Impact Factor and position within the hierarchy of evidence (Obremskey et al., 2005; Lau & Samman, 2007). However, a contemporaneous study of anaesthesia journals failed to find any statistically significant correlation between journal rank and evidence-based medicine principles (Bain & Myles, 2005). The variation encountered among scientific journals is further revealed by a study of seven medical/psychological journals which found highly varying adherence to statistical guidelines, irrespective of journal rank (Tressoldi et al., 2013). Analysis of statistical power in neuroscience studies (Button et al., 2013) found no significant correlation between statistical power and journal rank. The overall pattern from these studies suggests that journal rank is a poor proxy for methodological quality.

Colquhoun (2003) recounts his experience from publishing in Nature (with a then Impact Factor of 27.9) and only being cited 57 times, while another work published in a much lower impact journal (3.1) attained more than 400 citations! Clearly publication in a high impact journal does not guarantee that a paper will achieve the “much-desired dream of the author: to be read, cited, and remembered” (Albuquerque 2010). In our own example the main report of a two-year project, published within a monograph series with an Impact Factor of over 4.0, received only 18 citations whereas a methodological paper in a journal with an Impact Factor of only 2.37, a by-product from team musings during the main project, enjoyed 264 citations in the same time interval.

Traditional metrics conflate factors that are conceptually different, for example the popularity of an article and its intellectual contribution. This is analogous to ranking the performance of a sports team by the average number of spectators that view the team's matches. Although some studies demonstrate an association between quality and number of citations a significant number of studies fail to confirm this association. West and McIlwaine (2002) examined the association between peer ratings of quality and numbers of citations between 1997 and 2000 to articles appearing in the journal Addiction in 1997. Although two independent reviewers agreed moderately in their ratings of the papers, the correlation between these ratings and the number of citations was almost zero. More alarmingly one factor that was correlated with citation count was the region of origin of the first author of the paper. Noticeably papers from English speaking countries received more citations than those from continental Europe. These in turn received more than papers from the rest of the world. The reader will note that subjective evaluation of papers, using a range of unspecified cues, is here being used as a comparator to objective evaluation. The two raters may be consistently "wrong" or, more likely, their shared perception of quality relates to intrinsic qualities that are imperfectly captured by more objective measures. This study illustrates the associated challenges of measuring quality and of devising a methodology by which to demonstrate such associations.

Why people cite

Key to confusion surrounding use of metrics is the variety of reasons why people cite other authors. The worst offence an academic can perpetuate against a fellow academic is not misquoting or wrongly citing them but ignoring their work completely! Not only do numerous motives (psychological, sociological, political, historical, etc.) influence an author's decision to cite a study but the relative influence of these motives is likely to vary from discipline to discipline. Co-citation, rather than measuring scientific quality may, in fact, more accurately document the existence of a common paradigm and/or a community of interest (Simkin & Roychowdhury, 2003).

Further complexity derives from the finding that many authors categorically do not read the papers they cite from (Simkin & Roychowdhury, 2003). These authors estimate that only 20% of authors have read the work they cite. It is unclear whether this statistic will improve with greater open access to citeable sources or whether, conversely, it will get worse as it becomes easier to identify relevant work through World Wide Web search engines and social bookmarking.

Criticisms of established metrics

Numerous criticisms of the impact factor reveal that it falls short of established requirements for validity. For example the impact factor is not completely transparent or consistently reproducible (Rossner, 2007). It is potentially manipulable. However many criticisms stem not from questionable validity but more from inappropriate use (EASE, 2007).

Inappropriate use of the Impact Factor includes comparison across disciplines. Citation rates, distributions and patterns are highly diverse. Such variation can be detected across broad categories, for example when comparing mathematical sciences with biological sciences. However significant variation may pertain within

disciplines as when comparing medicine with dentistry, nursing or public health. Some degree of subject normalisation is required. However, deciding the boundaries for the field to be normalised is logistically and conceptually challenging.

Mathematically the assumptions underlying the impact factor are equally open to debate. Put simply the impact factor is an arithmetic mean. Statistics 101 affirms that an arithmetic mean is inappropriate as a metric for data with a skewed distribution (Joint Committee, 2008). To illustrate, about 90% of Nature's 2004 impact factor was based on only a quarter of its publications. Should we be attributing a metric to a journal where about three quarters of its publications perform worse than “average” (Nature, 2005)?

Technically concern has been expressed, with the advent of digital publications, that the “purity” of the Impact Factor as a measure has been diluted by so many confounding factors (for example digital only, digital and print, early view, payment for “gold access” etcetera) (Lozano et al, 2012). The potential for manipulation is further increased by the fact that digital journals without a print equivalent are not bounded by restrictions on the number of articles that they can accept.

When journal prestige is the battleground such concerns are significant enough. However debates regarding appropriateness are further intensified when the Impact Factor is proposed as a measure by which to evaluate institutional performance, as has been the case within the UK. Reworking the debate to assessment of the quality of individual articles, not the reputation of the journal in which they are published, has alleviated concerns. However, given the short timespan within which panel assessors assign each verdict, suspicion remains regarding the influence of a journal’s Impact Factor as a proxy for article quality.

Notwithstanding ongoing misgivings regarding the inappropriate use of Impact Factors as a single metric for journal or article excellence, progress has been made. More assessors are acknowledging that the Impact Factor should definitely not be used to assess individual researchers or institutions (Seglen 1997, EASE, 2007). This emerging consensus is concisely summarised in the EASE Statement (November 2007) which recommends that "journal impact factors are used only—and cautiously—for measuring and comparing the influence of entire journals, but not for the assessment of single papers, and certainly not for the assessment of researchers or research programmes". Logically such caution extends to other bibliometric measures examined at an individual level. The German Research Foundation (DFG) has published guidelines to evaluate only articles and not bibliometric information on candidates in all decisions concerning "performance-based funding allocations, postdoctoral qualifications, appointments, or reviewing funding proposals, [where] increasing importance has been given to numerical indicators such as the h-index and the impact factor" (DFG Press Release, 2010). Other influential bodies such as the National Science Foundation (US) and the Research Assessment Exercise (UK) have taken a similar stance.

In the UK, a parallel move to renewed interest in bibliometric measures has seen the opening up of the “impact agenda”. This movement mirrors a utilitarian focus on the usefulness of research to society – a philosophical stance guaranteed to alienate those preoccupied with pure science along with many within the arts and humanities research communities. Interestingly, media impact – i.e. “newsworthiness” - is excluded from a prodigious list of evidence sources for impact. Methodologically this recognises that newsworthiness is particularly vulnerable to manipulation by the author and institution. Newsworthiness also encapsulates the subjective judgements of editors and editorial staff on what interests a journalist audience (Chapman et al, 2007). This stance may extend to future treatment of altmetric phenomena such as Twitter, particularly given that a researcher need spend less time and money in stimulating interest by tweeting details of their research when compared to traditional routes such as crafting a press release. However newsworthiness clearly contributes to numbers of citations and so cannot be entirely removed from the picture.

Manipulation and game playing

How Individuals could manipulate metrics

Much alarm has been expressed at the prevalence of self-citation and its impact on citation metrics. While gratuitous self-citation is rightly condemned, as with all unashamed self publicity, a researcher specialising in a narrow field frequently has legitimate cause to cite their own contributions. Similarly although citation clubs or networks (“you cite my paper and I’ll cite yours”) are open to abuse what is more appropriate than a group of related researchers, who form a “virtual college”, citing each other’s relevant work? Arguably researchers perform a valuable educational service by drawing a reader’s attention to associated papers within the field. Indeed, the reverse argument, i.e. that in neglecting to cite relevant papers an author might be scientifically negligent, now holds increasing weight, particularly in the context of systematic reviews and the prevention of scientific waste (i.e. the commissioning of duplicate and redundant studies or the invisibility of research that is consequently underutilised in practice).

All disciplines can identify types of articles that may be widely cited, although not necessarily widely read. Basic laboratory methods, methodology texts, seminal reports of methods and publication standards are all potential candidates. For example the CONSORT statement for reporting randomised controlled trials, synchronously published in 2001 in several key journals, has attracted Google Scholar citations of over three thousand two hundred (twice – in BMC Medical Research Methodology; (Annals of Internal Medicine)), two thousand eight hundred (the Lancet) and two and a half thousand (BMJ) in its top four publishing channels. In this unique natural experiment the open access journal BMC Medical Research Methodology matched the Lancet even though only in its first year of publication.

How research teams/institutions could manipulate metrics

Many routes for individual manipulation are equally open to manipulation by research teams. While coordinated team self-citation may increase the scale of abuse it may also increase the likelihood of detection and censure. The same mechanisms of peer review and esteem that traditionally protect the academic citadel also defend against widescale abuse and manipulation of traditional metrics. Instances of academic fraud and large scale plagiarism are most frequently unearthed by fellow scientists. By implication serendipitous discovery of citation manipulation is more likely than

systematic identification of abuse. Manipulation of citations within a large research group or institution requires the determined collusion of a few influential individuals or the widespread complicity of a larger team. Nevertheless the rewards from academia, in terms of securing grants and tenure, are such that the traditional mechanisms may not be sufficient to deter systematic abuse.

How journals could manipulate metrics

While many express misgivings about the potential for manipulation and game playing, or indeed its prevalence, identified instances take the form of experiments to explore "what…if" scenarios. So, for example the specialist journal Folia Phoniatrica et Logopaedica sought to demonstrate the vulnerability of the Journal Impact Factor by publishing editorial content in 2007 that cited every single published article from that same journal over the previous two years (Opatrný, 2008). As a consequence the journal's Impact Factor increased from 0.66 to 1.44 until the Institute for Scientific Information suspended that journal's rating. More recently there are reported instances of "citation stacking" i.e. collusion of journals in citing each other with over twenty journals being suspended from ISI ratings either for this practice of uncontrolled self-citation.

A journal can adopt editorial policies to increase its impact factor.(Monastersky 2005; Arnold et al, 2011). For example, journals may publish a larger percentage of review articles which generally are cited more than research reports (Garfield, 1994). Thus review articles raise the impact factor of the journal and review journals often have the highest impact factors in their respective fields (Moustafa, 2014). Some journal editors set their submissions policy to "by invitation only" to invite senior scientists to publish "citable" papers to increase the journal impact factor (Moustafa, 2014).

Journals may attempt to limit the number of "citable items"—i.e., the denominator of the impact factor equation—by declining to publish articles (e.g. case reports in medical journals) that are unlikely to be cited or by altering articles (by not allowing an abstract or bibliography) in hopes that Thomson Scientific will not deem it a "citable item". As a result of negotiations over whether items are "citable", impact factor variations of more than 300% have been observed (PLOS Medicine Editors, 2006). Interestingly, items considered uncitable—and thus not incorporated in impact factor calculations—can, if cited, still enter the numerator part of the equation despite the ease with which such citations could be excluded. This effect is difficult to evaluate as the distinction between editorial comment and short original articles is not always obvious. For example, letters to the editor may refer to either class.

Strategically a journal might publish a large portion of its papers, or at least those expected to be highly cited, early in the calendar year. This gives papers the maximum time to gather citations. Several methods, not necessarily with nefarious intent, exist for a journal to cite articles in the same journal which will increase the journal's impact factor (Agrawal, 2005; Fassoulaki et al, 2002).

Coercive citation is a practice in which an editor forces an author to add citations to other articles from the same journal to an article before agreeing to publish it thereby inflating the journal's impact factor. A survey published in 2012 indicates that coercive citation has been experienced by one in five researchers working in economics, sociology, psychology, and business disciplines, and it is more common

in journals with a lower impact factor (Whilhite & Fong, 2012). Coercive citation has occasionally been reported for other scientific disciplines (Smith, 1997). Coercive citation approaches extortion and is perceived by many as a violation of scientific ethics.

While these examples occupy a whole spectrum of academic practice from legitimate exploitation of the rules through to questionable conduct and malpractice they collectively serve to illustrate the acknowledged weaknesses of existing metrics.

What have we learnt?

Assessing individual performance

Within the academic community there is growing recognition that traditional metrics must be used with caution when assessing individual performance. As Sahel (2011) observes:

> "Evaluating individual research performance is a complex task that ideally examines productivity, scientific impact, and research quality—a task that metrics alone have been unable to achieve"

At the same time the author acknowledges that evaluating individual scientific performance is an essential component of research assessment. Existing measures such as impact factor, numbers of citations and the "new indicators" (such as the h index and the g index) are all found wanting for such a task. It is relatively straightforward to examine the impact factor of the journals in which a particular person has published articles. This use is widespread, but controversial. Garfield (1998) warns of the "misuse in evaluating individuals" because there is "a wide variation from article to article within a single journal".

We do not conclude that more sophisticated indicators are required, to overcome long-acknowledged deficiencies such as the inability to discriminate between author order and the prevalence of cultural and language citation patterns. Instead we endorse a more encompassing range of indicators that factor in "teaching, mentoring, participation in collective tasks, and collaboration-building, in addition to quantitative parameters that are not measured by bibliometrics, such as number of patents, speaker invitations, international contracts, distinctions, and technology transfers". Clearly the potential of altmetrics comes not so much in improving the robustness of domains measured by traditional metrics but, more cogently, in expanding the range of activities catered for when assessing academic research performance.

Evaluating the performance of a research team, group, department or institution

An inherent attraction of bibliometrics in evaluating team or institutional performance is the apparent ease and speed with which assessments are performed, particularly compared to qualitative assessment by experts (Sahel, 2011). Haeffner-Cavaillon & Graillot-Gak (2009) describe their experience when evaluating the team performance of 600 research teams within the French INSERM research institution. They confirm that "analysis of bibliometric indicators cannot depend on one bibliometric indicator alone but must take into account several indicators to allow having an overall picture of the team output". They conclude that each indicator has its advantages and its

limitations, with caution required in not considering any single metric as an “absolute” index of scientific quality. They advocate a model of “enrichment” i.e. that metrics inform, yet not determine, a scientific committee’s debates. Furthermore they observe that, despite acknowledged limitations of peer review, most scientists appear to believe such a qualitative assessment is “the best system and agree that it is the only way to evaluate them”. Clearly qualitative assessment, informed but not determined by bibliometric indicators and supplemented by a wide range of altmetrics, will persist for some considerable time as the method of preference for assessment of team or institutional performance.

Recognising the contribution of a journal

Systems employing journal rank have been criticised for being “not only technically obsolete, but also counter-productive and a potential threat to the scientific endeavour” (Brembs et al, 2013). Generally, with notable exceptions, the impact factor has achieved widespread acceptance as an indicator of overall journal quality. However this may be associated less with overall validity and more with its ubiquitous and highly visible presence. Debates about the appropriateness of the impact factor centre on the optimal unit for comparison. Comparisons at discipline, sub-discipline or topic level attract expressions of dissatisfaction, most typically from those who feel disadvantaged within a particular constituency or configuration. Academics recognise that the distribution of numbers of citations for articles within a journal is heavily skewed with a small number of articles holding undue influence over the overall impact factor. Impact factors may preserve the status quo with a minimum two years before a journal may apply for an impact factor. Under such circumstances it may be challenging to lure manuscripts away from established journals within a self-preserving hierarchy. Interestingly, the role of the impact factor has evolved to be far more about signifying the prestige of the journal targeted by a paper pre-publication (i.e. the aspiration). Once a paper is published, the influence of the impact factor metric may be diluted by additional insights from individual article level metrics.

Methodologically the inability of the impact factor to distinguish open access from traditional subscription-based journals is a current cause for concern. However this may represent nothing more than a transitory blip in a relentless stampede towards open access. Harnad (2008) identifies collective contributors to “Open Access Impact Advantage”. These include:

- an early access advantage (with a preprint being accessible before the published postprint),
- a quality bias (higher quality articles are more likely to be made OA),
- a quality advantage (higher quality articles benefit more from being made OA for users who cannot otherwise afford access),
- a usage advantage (OA articles are more accessible, more quickly and easily, for downloading), and
- a competitive advantage (which will vanish once all articles are OA)

He concludes that open access possesses nett benefits for research and researchers across all disciplines.

A persisting challenge posed to altmetrics is the mixed economy within which journals continue to operate. At present activity by significant numbers of paper-based readers is not factored into electronic metrics. While underreporting is one consideration, of greater concern is the non-representative distribution of readers. For example those in developing countries may read articles in paper form and may encounter delays in receiving printed journals. However the economics of paper journals may challenge these stereotypes as publishers recognise that offering nominally priced access to readers from developing countries, with comparatively minimal investment in additional infrastructure, may extend the potential readership and result in a nett financial gain.

Some Alternative traditional metrics

Related indices

Alternative traditional indices include:

- Immediacy index: the number of citations the articles in a journal receive in a given year divided by the number of articles published
- Cited half-life: the median age of the articles cited in Journal Citation Reports each year. For example, if a journal's half-life in 2005 is 5, citations from 2001-2005 constitute half of all the citations from that journal in 2005 with the remaining citations preceding 2001.
- Aggregate impact factor for a subject category: calculated taking into account the number of citations to all journals in the subject category and the number of articles from all journals in the subject category
- Source normalized impact per paper (SNIP) is a factor released in 2012 by Elsevier to estimate impact (Elsevier, 2014). The measure is calculated as SNIP=RIP/(R/M), where RIP=raw impact per paper, R = citation potential and M = median database citation potential (Moed, 2010).

These measures apply only to journals, not individual articles or individual scientists, unlike the h-index. The relative number of citations an individual article receives is better viewed as citation impact.

A note about Altmetrics

While others in this book will offer expert evaluations of the future role of altmetrics I choose to showcase the “trad-alt paradox”. Alternative metrics (alt-metrics) typically measure impact at an article level. They include article views, downloads or mentions in social media, e.g. Twitter. As a leading player in developing an online journal “presence”, the BMJ published the number of views for its articles. These corresponded somewhat to patterns exhibited by citations (Delamothe & Smith, 2004). The paradox is this – if trad-metrics are widely considered as being flawed and of questionable accuracy then any such correlation should, in theory, be a cause for concern, not celebration. Additionally, if altmetrics add little to the existing picture then there is little justification for expending much time and effort in expanding their use. The converse is that if altmetrics rightly claim to add value over the use of trad-metrics we would expect them to capture a different perspective from that revealed by their traditional counterparts. If however, altmetrics, are substantively different from their predecessors then they risk losing credibility and being discredited. Resolution of this paradox may not come by attaining greater validity but may lie in

other advantages, e.g. the facility to derive an earlier picture of quality distinctions or to predict the eventual performance of an article via altmetrics.

A further irony is that the altmetric community seeks to establish credibility by mimicking its forebears. Thus in 2008 the well-regarded Jounal of Medical Internet Research began publishing metrics on both views and Tweets. On the basis of a reasonably good indication of highly cited articles Eysenbach (2011) proposed a "Twimpact factor" (number of Tweets an article receives in the first seven days of publication) and a “Twindex” (the rank percentile of an article's Twimpact factor). Although imitation may be the sincerest form of flattery altmetrics may prosper more from developing unique and creative ways of demonstrating impact rather than in slavishly imitating the terminology or methodology of traditional metrics.

Some tentative conclusions

Brembs et al (2013) conclude with the pessimistic verdict “Given the data we surveyed above [i.e. about journal ranking systems], almost anything appears superior to the status quo”. Harnad (2008) examines whether traditional metrics will to continue to have a role within a world increasingly populated by social media and social networks. To take just one example, namely download counts (Hitchcock et al 2003), these are rightly considered a metric of computer activity not scholarship. Downloads may not read. Their function may be analogous to download of music when compared to live streaming. Just as a single music track may be played multiple times without registering statistics, so too an article may experience multiple use without being captured by metrics. Furthermore multiple versions of the same article e.g. publisher and repository version, may cause counting anomalies analogous to those previously experienced by journals with both print and electronic versions of the same article.

Given the flaws associated with traditional metrics identified from the published research, along with often damning verdicts pronounced by informed critics of the metrics and by the academic community at large, it is tempting to sound a death knell for the measures of the immediate past. However as already stated, and as will become clear from the remainder of this book, the new generation of alt-metrics is neither a more accurate representation of academic “quality” nor is it immune from criticism.

Interesting results from research seek to isolate the respective value of traditional and alt-metrics. For example a study in International Journal of Cardiology compared a top 20 based on downloads from the journal with a top 20 of most cited articles from the same journal (Coats, 2005). There was no overlap between the two lists. Perneger (2004) studied a cohort of papers published in the BMJ in 1999, finding that the hit count on the website in the week after online publication predicted the number of citations in subsequent years A more recent study (Lokker et al, 2008) confirmed that the citation performance of journal articles can be predicted extremely early, even within three weeks of publication.

What can we conclude from the above? First, although the prevailing opinion is that use of metrics is intended to reflect quality we should not be oblivious to the fact that choice of metrics reflects how quality is contemporaneously perceived. A move towards increased use of simple download statistics would offer a strident statement

that popularity is one important arbiter of quality; a verdict open to immediate challenge from anyone who compares the Top 40 bestselling music tracks of the year with a selection for the same year from informed music critics. Second, lack of consistency between different measures, such as downloads and citations, is not a problem but an opportunity – an opportunity to reflect a more holistic and nuanced appreciation of what academic quality really means. Of course selection of which metrics are to be used for which purpose – for evaluating an individual, a team or a journal - becomes no less problematic, even if the improved ease and sophistication with which such data is collected moves us away from invidious either/or choice scenarios. Finally, in a book that seeks to address an imbalance between long-established traditional metrics and up-and-coming altmetrics, it is hopefully not too subversive to extend a plea to use both sets side by side in a display of, what I shall label "complemetrics"!

References

Agrawal, A. (2005) Corruption of Journal Impact Factors. Trends in Ecology and Evolution, **20** (4), 157.

Albuquerque, U. P. (2010) The tyranny of the impact factor: why do we still want to be subjugated?. Rodriguésia-Instituto de Pesquisas Jardim Botânico do Rio de Janeiro, **61**(3).

Arnold, D N. and Fowler, K K. (2011) Nefarious Numbers. Notices of the American Mathematical Society, **58** (3), 434–437.

Bain, C. R., and Myles, P. S. (2005) Relationship between journal impact factor and levels of evidence in anaesthesia. Anaesthesia and intensive care, **33**(5), 567-570.

Bence, V., and Oppenheim, C. (2005) The evolution of the UK's Research Assessment Exercise: publications, performance and perceptions. Journal of Educational Administration and History, **37**(2), 137-155.

Boyer, E. L. (1990) Scholarship reconsidered: Priorities of the professoriate. Princeton, NJ: Carnegie Foundation for the Advancement of Teaching.

Brembs, B., Button, K., and Munafò, M. (2013) Deep impact: unintended consequences of journal rank. Frontiers in Human Neuroscience, **7**, 291.

Button K. S., Ioannidis J. P. A., Mokrysz C., Nosek B. A., Flint J., Robinson E. S. J., et al. (2013) Power failure: why small sample size undermines the reliability of neuroscience. Nat. Rev. Neuroscience, **14**, 365–376

Callaham, M., Wears, R. L., & Weber, E. (2002) Journal prestige, publication bias, and other characteristics associated with citation of published studies in peer-reviewed journals. JAMA, **287**(21), 2847-2850.

Chapman, S., Nguyen, T. N., and White, C. (2007) Press-released papers are more downloaded and cited. Tobacco control, **16**(1), 71.

Coats, A. J. S. (2005) Top of the charts: download versus citations in the International Journal of Cardiology. International Journal of Cardiology, **105** (2), 123-125.

Colquhoun, D. (2003) Challenging the tyranny of impact factors. Nature, **423**(6939), 479.

de Lacey, G., Record, C., and Wade, J. (1985) How accurate are quotations and references in medical journals? Br Med J (Clin Res Ed), **291**(6499): 884-6.

Delamothe, T, and Smith, R (2004) Open access publishing takes off. BMJ, **328**:1-3.

Deutsche Forschungsgemeinschaft (DFG) German Research Foundation [Press release]

<http://www.dfg.de/en/service/press/press_releases/2010/pressemitteilung_nr_07/ind Retrieved 17/10/2014
Dimitrov, J. D., Kaveri, S. V., & Bayry, J. (2010) Metrics: journal's impact factor skewed by a single paper. Nature, **466**(7303), 179-179.
Dobell, C. (1938) Dr O Uplavici (1887-1938). Parasitology, **30**, 239-41.
EASE (2007) European Association of Science Editors (EASE) Statement on Inappropriate Use of Impact Factors. European Association of Science Editors. Reading.
Elsevier (2014) Elsevier Announces Enhanced Journal Metrics SNIP and SJR Now Available in Scopus. Press release. Elsevier. Retrieved 2014-07-27.
Eysenbach, G. (2011) Can tweets predict citations? Metrics of social impact based on Twitter and correlation with traditional metrics of scientific impact. J Med Internet Res, Dec 19; **13**(4), e123. doi: 10.2196/jmir.2012.
Fassoulaki, A., Papilas, K., Paraskeva, A., and Patris, K. (2002) Impact factor bias and proposed adjustments for its determination. Acta Anaesthesiologica Scandinavica, **46**(7), 902-905.
Garfield, E (20 June 1994) The Thomson Reuters Impact Factor. Thomson Reuters.
Garfield, E (June 1998). The Impact Factor and Using It Correctly. Der Unfallchirurg, **101** (6), 413–414.
Grant, B (21 June 2010) New impact factors yield surprises. The Scientist. Accessed at: http://www.the-scientist.com/?articles.view/articleNo/29093/title/New-impact-factors-yield-surprises/ Retrieved 17/10/2014.
Haeffner-Cavaillon, N. and Graillot-Gak, C. (2009) The use of bibliometric indicators to help peer-review assessment. Arch Immunol Ther Exp (Warsz), **57,** 33–38. doi: 10.1007/s00005-009-0004-2.
Harnad, S. (2008) Validating research performance metrics against peer rankings. Ethics in Science and Environmental Politics, **8**(11).
Harzing, A., & van der Wal, R. (2008) Google Scholar as a new source for citation analysis. Ethics in Science and Environmental Politics, **8**(1), 61–73.
Hitchcock, S; Woukeu, A; Brody, T; Carr, L; Hall, W and Harnad, S. (2003) Evaluating Citebase, an open access Web-based citation-ranked search and impact discovery service http://eprints.ecs.soton.ac.uk/8204/ Retrieved 17/10/2014
Hirsch, J E., (2005) An index to quantify an individual's scientific research output. Proceedings of the National Academy of Sciences, **102**(46) 16569-16572 http://www.pnas.org/cgi/content/abstract/102/46/16569 Retrieved 17/10/2014
Joint Committee on Quantitative Assessment of Research (12 June 2008).Citation Statistics (PDF). International Mathematical Union. Retrieved 17/10/2014
Kjaergard, L. L., and Gluud, C. (2002) Citation bias of hepato-biliary randomized clinical trials. Journal of Clinical Epidemiology, **55**(4), 407-410.
Lau, S. L., and Samman, N. (2007) Levels of evidence and journal impact factor in oral and maxillofacial surgery. International journal of oral and maxillofacial surgery, **36** (1), 1-5.
Lokker, C., McKibbon, K., McKinlay, R. J., Wilczynski, N. L., & Haynes, R. B. (2008) Prediction of citation counts for clinical articles at two years using data available within three weeks of publication: retrospective cohort study. BMJ, **336**(7645), 655-657.
Lozano, G. A.; Larivière, V; Gingras, Y (2012) The weakening relationship between the impact factor and papers' citations in the digital age. Journal of the American Society for Information Science and Technology, **63** (11), 2140. .

MacRoberts, M. H., & MacRoberts, B. R. (1989) Problems of citation analysis: A critical review. Journal of the American Society for Information Science, **40**(5), 342-349.
Moed, H (2010) Measuring contextual citation impact of scientific journals. Journal of Informetrics, **4**, 256–277.
Monastersky, R (14 October 2005) The Number That's Devouring Science. The Chronicle of Higher Education, **52**(8),14.
Moustafa, K (2014) The disaster of the impact factor. Science and Engineering Ethics, **21** (1), 139-142.
Nature (2005) Not-so-deep impact. Nature **435** (7045), 1003–1004. doi: 10.1038/4351003b.
Obremskey W. T., Pappas N., Attallah-Wasif, E., Tornetta P., and Bhandari, M. (2005) Level of evidence in orthopaedic journals. J. Bone Joint Surg. Am, **87**, 2632–2638.
Opatrný, T. (2008) Playing the system to give low-impact journal more clout. Nature, **455** (7210), 167-167.
Perneger, T. V (2004) Relation between online "hit counts" and subsequent citations: Prospective study of research papers in the BMJ. BMJ, **329** (7465): 546–7.
PLoS Medicine Editors (6 June 2006) The Impact Factor Game. PLoS Medicine **3** (6), e291. .
Reedijk, J. (1998) Sense and nonsense of science citation analyses: comments on the monopoly position of ISI and citation inaccuracies. Risks of possible misuse and biased citation and impact data. New Journal of Chemistry, **22**(8), 767-770.
Rossner, M., Van Epps, H., and Hill, E. (2007) Show me the data. Journal of Cell Biology, **179** (6), 1091–2
Sahel, J. A. (2011) Quality versus quantity: assessing individual research performance. Sci Transl Med, May 25; **3**(84), 84cm13.
Schoonbaert, D., and Roelants, G. (1996) Citation analysis for measuring the value of scientific publications: quality assessment tool or comedy of errors? Trop Med Int Health, Dec;**1**(6), 739-52.
Seglen, P. O. (1997) Why the impact factor of journals should not be used for evaluating research. BMJ, **314**(7079), 497.
Simkin, M. V. and Roychowdhury, V. P. (2003) Read before you cite! Complex Systems, **14,** 269-274
Smith, R. (1997) Journal accused of manipulating impact factor. BMJ, **314**(7079), 461.
Tressoldi, P. E., Giofré, D., Sella, F., and Cumming, G. (2013) High impact = high statistical standards? Not necessarily so. PloS One, **8**(2), e56180.
van Raan, A. F. (2003) The use of bibliometric analysis in research performance assessment and monitoring of interdisciplinary scientific developments. Technikfolgenabschätzung, **12**(1), 20–29. English translation available: http://www.itas.fzk.de/tatup/031/raan03a.htm
van Raan, A. J. F. (2005) Fatal attraction: Conceptual and methodological issues problems in the ranking of universities by bibliometric methods. Scientometrics, **62**(1), 133-143.
West, R., & McIlwaine, A. (2002) What do citation counts count for in the field of addiction? An empirical evaluation of citation counts and their link with peer ratings of quality. Addiction, **97**(5), 501-504.
Wilhite, A. W. and Fong, E. A. (2012). Coercive Citation in Academic Publishing. Science, **335** (6068), 542–3.