Validating research performance metrics against peer rankings

Stevan Harnad*

Chaire de recherche du Canada, Institut des sciences cognitives, Université du Québec à Montréal, Montréal, Québec H3C 3P8, Canada

Department of Electronics and Computer Science, University of Southampton, Highfield, Southampton SO17 1BJ, UK

ABSTRACT: A rich and diverse set of potential bibliometric and scientometric predictors of research performance quality and importance are emerging today — from the classic metrics (publication counts, journal impact factors and individual article/author citation counts) to promising new online metrics such as download counts, hub/authority scores and growth/decay chronometrics. In and of themselves, however, metrics are circular: They 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; a unique opportunity to do this is offered by the 2008 UK Research Assessment Exercise, in which a full spectrum of metrics can be jointly tested, field by field, against peer rankings.

KEY WORDS: Bibliometrics · Citation analysis · Journal impact factor · Metric validation · Multiple regression · Peer review · Research assessment · Scientometrics · Web metrics

INTRODUCTION

Philosophers have a saying[1] (about those who are sceptical about metaphysics): 'Show me someone who wishes to refute metaphysics and I'll show you a metaphysician with a rival system' (meaning that there is no escaping metaphysics one way or the other: even anti-metaphysics is metaphysics). The same could be said of bibliometrics, or more broadly scientometrics.

If we divide the evaluation of scientific and scholarly research into (1) subjective evaluation (peer review) and (2) objective evaluation (scientometrics: henceforth 'metrics'), then even those who wish to refute metrics in favor of peer review first have to demonstrate that peer review (Harnad 2004a) is somehow more reliable and valid than metrics: moreover, in order to demonstrate that, in turn, without circularity (i.e. without simply decreeing that peer review is better because peers agree on what research is better and they also agree that peer review is better than metrics!), peer review, too, will have to be evaluated objectively (i.e. via metrics).

This is not to say that metrics themselves are exempt from the need for validation. However, trying to validate unvalidated metrics against unvalidated metrics is no better than trying to validate peer review with peer review: Circularity has to be eliminated on both sides.

Other contributions to this Theme Section have done a good job pointing out the inappropriateness of the use of unvalidated journal impact factors (JIFs) for evaluating anything, be it journal quality, research quality, or researcher quality (e.g. Campbell 2008). Not only is the JIF, in and of itself, not validated as a

_____

[1] In 'Appearance and Reality', Bradley (1897/2002) wrote (of Ayer) that the man who is ready to prove that metaphysics is wholly impossible … is a brother metaphysician with a rival theory.

*Email: harnad@ecs.soton.ac.uk

measure of journal quality, especially when comparing across different fields, but (being a journal average) it is a particularly blunt instrument for evaluating and comparing individual authors or papers. Comparing authors in terms of their JIFs is like comparing university student applicants in terms of the average marks of the secondary schools from which they graduated, instead of comparing them in terms of their own individual marks (Moed 2005).

VALIDATING METRICS

Psychometrics of cognitive performance capacity

Even author citation counts stand unvalidated in and of themselves. The problem can be best illustrated with an example from another metric field: psychometrics (Kline 2000). If we wish to construct a test of human aptitude, it is not sufficient simply to invent test items that we hypothesize to be measuring the performance capacity in question, and use those items to construct a set that is internally consistent (i.e. higher scorers tend to score higher on all items, and vice versa) and repeatable (i.e. the same individual tends to get the same score on repeated sittings). So far, this is merely a *reliable* test, not necessarily a *valid* one.

Let us call the capacity we are trying to measure and predict with our test our 'criterion.' To validate a psychometric test, we have to show that either (1) the test has *face-validity* (i.e. that it is a direct measure of the criterion, as in the case of a long-distance swimming test to test long-distance swimming ability, or a calculational test to test calculating ability) or (2) our test is strongly correlated with a face-valid test of the criterion or with a test that has already been validated (as being correlated with the criterion).

Scientometrics of research performance quality

In psychometrics, it is the correlation with the criterion that gets us out of the problem of circularity. But what is the criterion in the case of scientometrics? Presumably it is research performance quality itself. But what is the face-valid measure of research performance quality? Apart from the rare cases where a piece of research instantly generates acknowledged break-throughs or applications, the research cycle is too slow and uncertain to provide an immediate face-valid indicator of quality. So what do we do? We turn to expert judgment: journals (and research funders) consult qualified peer referees to evaluate the quality of research output (or, in the case of grants, the quality of research proposals).

Now, as noted, peer review itself stands in need of validation, just as metrics do. Even if we finesse the problem of reliability, by only considering peer judgments on which there is substantial agreement (Harnad 1985), it still cannot be said that peer review is a face-valid measure of research quality or importance, just as citation counts are not a face-valid measure of the same.

Getting metrics off the ground

It is useful again to return to the analogous case of psychometrics: How did IQ testing first get off the ground, given that there was no face-valid measure of intelligence? IQ tests were bootstrapped in 2 ways: First, there were 'expert' ratings of pupils' performance by their teachers. Teacher ratings are better than nothing, but of course they too, like peer review, are neither face-valid nor already validated. Second, there was the reasonable hypothesis that, whatever intelligence was, the children who at a given age could do what most children could only do at an older age were more likely to be more intelligent (and vice versa). IQ refers to the 'Intelligence Quotient': the ratio of an individual child's test scores (mental age) to the test norms for their own age (chronological age). Now, this risks being merely a measure of precociousness or developmental delay, rather than intelligence, unless it can be shown that in the long run the children with the higher IQ ratios do indeed turn out to be the more intelligent ones. Psychometricians had the advantage of being able to follow children and their test scores and their teacher ratings through their life cycles long enough and on a large enough population to be able to validate and calibrate the tests they constructed against their later academic and professional performance. Once tests are validated, the rest becomes a matter of optimization through calibration and fine-tuning, including the addition of further tests.

Multiple metrics: multiple regression

Psychometric tests and performance capacity turn out to be multifactorial: no single test covers all of our aptitudes. It requires a battery of different tests (e.g. of reasoning ability, calculation, verbal skill and spatial visualization) to be able to make an accurate assessment of individuals' performance capacity and to predict their future academic and professional success. There exist general cognitive abilities as well as domain-specific special abilities (such as those required for music, drawing or sports). Even the domain-general abilities (such as reasoning or verbal

comprehension) can be factored into a large single general intelligence factor, or 'G', plus a number of lesser cognitive factors (Kline 2000). Each test has differential weightings on the underlying factors, and that is why multiple tests rather than a single test need to be used for evaluation and prediction.

Scientometric measures do not consist of multiple tests with multiple items (Moed 2005). They are individual 1-dimensional metrics, such as journal impact factors or individual citation counts. Some *a priori* functions of several variables such as the h-index (Hirsch 2005) have also been proposed recently, but they too yield 1-dimensional metrics. Many further metrics have been proposed or are possible, among them (1) download counts (Hitchcock et al. 2003), (2) chronometrics (growth- and decay-rate parameters for citations and downloads; Brody et al. 2006), (3) Google PageRank-like recursively weighted citation counts (citations from highly cited articles or authors get higher weights; Page et al. 1999), (4) co-citation analysis, (5) hub/authority metrics (Kleinberg 1999), (6) endogamy/exogamy metrics (narrowness/width of citations across co-authors, authors and fields), (7) text-overlap and other semiometric measures, (8) prior research funding levels, doctoral student counts, and other nonbibliometric performance indicators (Harnad 2004b, Harzing 2008 [this Theme Section]).

Without exception, however, none of these metrics can be said to have face validity: They still require objective validation. How to validate them? Jointly analyzing them for their intercorrelational structure could yield some common underlying factors that each metric measures to varying degrees, but that would still be circular because neither the metrics nor the factors have been validated against their external criterion.

Validating metrics against peer rankings

What is that external criterion — the counterpart of psychometric performance capacity — in the case of research performance quality? The natural candidate is peer review. Peer review does not have face-validity either, but we rely on it already, and it is what critics of metrics typically recommend in place of metrics. So, the natural way to test the validity of metrics is against peer review. If metrics and per rankings turn out to be uncorrelated, that will be bad news. If they turn out to be strongly correlated, then we can have confidence in going on to use the metrics independently. Peer rankings can even be used to calibrate and optimize the relative 'weights' on each of the metrics in our joint battery of candidate metrics, discipline by discipline.

The simplest case of linear regression analysis is the correlation of one variable (the 'predictor') with another (the 'criterion'). Correlations can vary from +1 to –1. The square of the correlation coefficient indicates the percentage of the variability in the criterion that is predictable from the predictor. In multiple regression analysis, there can be P different predictors and C different criteria. Again, the square of the overall PC correlation indicates what percentage of the variability in the criteria is jointly predictable from the predictors. Each of the individual predictors also has a ('beta') weight that indicates what proportion of that overall predictability is contributed by that particular variable.

If we take peer review rankings as our (single) criterion (having first tested multiple peer rankings for reliability), and we take our battery of candidate metrics as our predictors, this yields a multiple regression equation of the form $b_1P_1 + b_2P_2 + \ldots + b_pP_p = C$. If the overall correlation of P with C is high, then we have a set of metrics that has been jointly validated against peer review (and, incidentally, vice versa). The metrics will have to be validated separately field by field, and their profile of beta weights will differ from field to field. Even after validation, the initialized beta weights of the battery of metrics for each research field will still have to be calibrated, updated and optimized, in continuing periodic cross-checks against peer review, along with ongoing checks on internal consistency for both the metrics and the peer rankings. However, the metrics will have been validated.

The Research Assessment Exercise in the UK

Is there any way this validation could actually be done? After all, journal peer reviews (as well as grant-proposal peer review) are done piece-wise, locally, and their referee ratings are both confidential and un-normalized. Hence, they would not be jointly useable and comparable even if we had them available for every paper published within each field. There is, however, one systematic database that provides peer rankings for all research output in all fields at the scale of the entire research output of a large nation and research provider: The UK's Research Assessment Exercise (RAE) (Harnad 2007, Butler 2008 [this Theme Section]).

For over 2 decades, the UK assembled peer panels to evaluate and rank the research output of every active researcher in every department of every UK university every 6 yr. After each evaluation, the departments were accorded top-sliced research funding in proportion to their RAE ranks. The process was very costly and time-consuming. Moreover, it has been shown in a number of correlational studies that the peer rankings are highly correlated with citation metrics in all fields tested (Oppenheim 1996, Holmes & Oppenheim 2001,
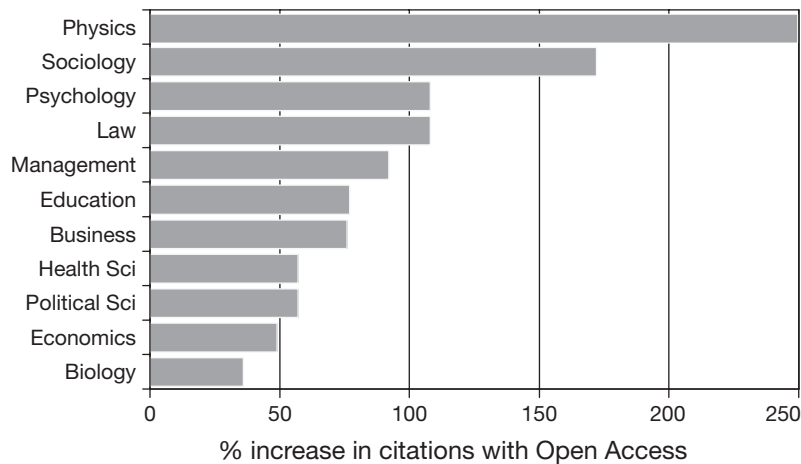
Fig. 1. Percent increase in citations for articles (in the same issue and journal) that are made freely accessible online (Open Access, OA) compared to those that are not. The OA advantage has been found in all fields tested. (Data from Harnad & Brody 2004 and Hajjem et al. 2005)

Smith & Eysenck 2002, Norris & Oppenheim 2003), even though citations were not counted in doing the peer rankings. It was accordingly decided that after one grand parallel ranking/ metrics exercise in 2008, the RAE would be replaced by metrics alone (supplemented by 'light-touch' peer review in some fields).

The open access research web: a synergy

The database for the last RAE (2008) hence provides a unique opportunity to validate a rich and diverse battery of candidate metrics for each discipline. The broader the spectrum of potential metrics tested, the greater the potential for validity, predictiveness, and customizability according to each discipline's own unique profile. As a bonus, generating and harvesting metrics on the open access research web will not only help measure and predict research performance and productivity, but will also help maximize it (Shadbolt et al. 2006).

It has now been demonstrated in over a dozen disciplines, systematically comparing articles published in the same journal and year, that the citation counts of articles that are made freely accessible to all users on the web (Open Access, OA) are on average twice as high as the citation counts of those that are not (Lawrence 2001, Harnad & Brody 2004, Hajjem et al. 2005; see Fig. 1).

There are many different factors contributing to this 'OA impact advantage', including an *early access advantage* (when the preprint is made 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 down-loading) and a *competitive advantage* (which will vanish once all articles are OA). It is clear that OA is a net benefit to research and researchers in all fields.

Just as peer rankings and metrics can be used to mutually validate one another, so metrics can be used as incentives for providing OA, while OA itself (as it grows) enhances the predictive and directive power of metrics (Brody et al. 2007). The prospect of increasing both their usage metric and their citation metrics (and their attendant rewards) is an incentive to researchers to provide OA to their findings. The resulting increase in openly accessible research not only means more research access, usage and progress, but it provides more open ways to harvest, data-mine and analyze both the research findings and the metrics themselves. This means richer metrics, and faster and more direct feedback between research output and metrics, helping to identify and reward ongoing research, and even to help set the direction for future research.

A foretaste of the open access research web is given by Citebase, a scientometric search engine (Hitchcock et al. 2003, Brody et al. 2006). Drawing mostly on the Physics Arxiv, Citebase reference-links nearly 500 000 papers and ranks papers and authors based on citation counts, download counts, and various other metrics (see www.citebase.org/help/order). It also generates growth curves for downloads and citations (see www.citebase.org/abstract?id=oai%3AarXiv.org%3A hep-th%2F0012054). Early download growth predicts later citation growth (Brody et al. 2006). Currently, ranking can only be done one metric at a time, but Citebase can be redesigned to rank using multiple metrics jointly, and even to adjust the weight (from –1 to +1) on each metric. This could be used to calibrate the outcome of the multiple regression analysis described earlier for validating metrics. Exploratory analysis as well as fine-tuning adjustments could then be done by calibrating the beta weights.

LITERATURE CITED

Bradley FH (1897/2002) Appearance and reality: a metaphysical essay. Adamant Media Corporation, Boston, MA

Brody T, Harnad S, Carr L (2006) Earlier web usage statistics as predictors of later citation impact. J Am Soc Information Sci Technol (JASIST) 57:1060–1072. http://eprints.ecs. soton.ac.uk/10713/

Brody T, Carr L, Gingras Y, Hajjem C, Harnad S, Swan A (2007) Incentivizing the open access research web: publication-archiving, data-archiving and scientometrics. CTWatch Quarterly 3(3). http://eprints.ecs.soton.ac.uk/ 14418/

Butler L (2008) Using a balanced approach to bibliometrics: quantitative performance measures in the Australian Research Quality Framework. Ethics Sci Environ Polit 8: 83–92

Campbell P (2008) Escape from the impact factor. Ethics Sci Environ Polit 8:5–7

Hajjem C, Harnad S, Gingras Y (2005) Ten-year cross-disciplinary comparison of the growth of open access and how it increases research citation impact. IEEE Data Eng Bull 28(4):39–47. http://eprints.ecs.soton.ac.uk/ 11688/

Harnad S (1985) Rational disagreement in peer review. Sci Technol Human Values 10:55–62. http://cogprints.org/ 2128/

Harnad S (2004a) The invisible hand of peer review. In: Shatz B (ed) Peer review: a critical inquiry. Rowland & Littlefield, Lanham, MD, p 235–242. http://cogprints. org/1646/

Harnad S (2004b) Enrich impact measures through open access analysis. BMJ 2004:329. http://bmj.bmjjournals. com/cgi/eletters/329/7471/0-h#80657

Harnad S (2007) Open access scientometrics and the UK research assessment exercise. In: Torres-Salinas D, Moed HF (eds) Proc 11th Annu Meet Int Soc Scientometrics and Informetrics, 25–27 Jun 2007 Madrid, p 27–33. http://eprints.ecs.soton.ac.uk/13804/

Harnad S, Brody T (2004) Comparing the impact of open access (OA) vs. non-OA articles in the same journals. D-Lib 10. www.dlib.org/dlib/june04/harnad/06harnad.html

Harzing AWK, van der Wal R (2008) Google Scholar as a new source for citation analysis. Ethics Sci Environ Polit 8:61–73

Hirsch JE (2005) An index to quantify an individual's scientific research output. Proc Natl Acad Sci 102:16569–16572. www.pnas.org/cgi/content/abstract/102/46/16569

Hitchcock S, Woukeu A, Brody T, Carr L, Hall W, Harnad S (2003) Evaluating Citebase, an open access web-based citation-ranked search and impact discovery service. http://eprints.ecs.soton.ac.uk/8204/

Holmes A, Oppenheimer C (2001) Use of citation analysis to predict the outcome of the 2001 Research Assessment Exercise for Unit of Assessment (UoA) 61: Library and Information Management. Information Research 6(2). http://informationr.net/ir/6-2/paper103.html

Kleinberg J M (1999) Hubs, authorities, and communities. ACM Computing Surveys 31(4). www.cs.brown.edu/ memex/ACM_HypertextTestbed/papers/10.html

Kline P (2000) The new psychometrics: science, psychology and measurement. Routledge, London

Lawrence S (2001) Online or invisible? Nature 411:521. http://citeseer.ist.psu.edu/online-nature01/

Moed HF (2005) Citation analysis in research evaluation. Springer, New York

Norris M, Oppenheim C (2003) Citation counts and the Research Assessment Exercise V: archaeology and the 2001 RAE. J Documentation 59(6)709–730. www.garfield. library.upenn.edu/papers/oppenheim.pdf

Oppenheim C (1996) Do citations count? Citation indexing and the research assessment exercise. Serials 9:155–161. http://uksg.metapress.com/index/5YCDB0M2K3XGAYA6.pdf

Page L, Brin S, Motwani R, Winograd T (1999) The PageRank citation ranking: bringing order to the web. http://dbpubs. stanford.edu:8090/pub/1999–66

Shadbolt N, Brody T, Carr L, Harnad S (2006) The open research web: a preview of the optimal and the inevitable. In: Jacobs N (ed) Open access: key strategic, technical and economic aspects. Chandos, London. http://eprints.ecs. soton.ac.uk/12453/

Smith AT, Eysenck M (2002) The correlation between RAE ratings and citation counts in psychology. Tech Rep. http://cogprint.org/2749/