Harnad, S. (submitted, 2008) Validating Research Performance Metrics Against Peer Rankings. *Inter-Research Ethics in Science and Environmental Politics*. Theme Section on: 'The use and misuse of bibliometric indices in evaluating scholarly performance'

Validating Research Performance Metrics Against Peer Rankings

Stevan Harnad
Chaire de recherche du Canada
Institut des sciences cognitives
Universite du Quebec a Montreal
Montreal, Quebec, Canada H3C 3P8
and
Department of Electronics and Computer Science
University of Southampton
Highfield, Southampton
SO17 1BJ UNITED KINGDOM
http://www.ecs.soton.ac.uk/~harnad/

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, and 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 just "metrics"), then even those who wish to refute metrics in favor of peer review first have to demonstrate that peer review (2004a) is somehow more reliable and valid than metrics: And to demonstrate that 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 from validation either. 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.

The other contributions to this special ESEP special issue have done a good job pointing out the inappropriateness of the unvalidated use of journal impact factors (JIFs) in evaluating anything, be it journal quality, research quality, or researcher quality (Campbell 2008). Not only is the JIF, in and of itself, not validated as a 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 the applicants have graduated, instead of comparing them in terms of their own individual marks (Moed 2005).

VALIDATING METRICS

Psychometrics of Cognitive Performance Capacty

But 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, that 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 either that the test has *face-validity* (i.e., that it is itself 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, in the absence of face-validity, we have to show that our test is strongly correlated either 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 research quality or importance.

Getting Metrics of 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 two ways: First, there were (1) "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.

In addition, there was the reasonable hypothesis that, whatever intelligence was, (2) 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) . The "Q" in 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. And in that case 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 turned out to be multifactorial: No single test covers all of our aptitudes. It requires a battery of different tests (of reasoning ability, calculation, verbal skill, spatial visualization, etc.) 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, sports); and even the domain-general abilities 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 just 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 one-dimensional metrics, such as journal impact factors or individual citation counts. Some apriori functions of several variables such as the h-index (Hirsch 2005) have also been proposed

recently, but they too yield one-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, etc. (Harnad 2004b; Harzing 2008).

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 (a) we rely on it already and (b) 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 variable that is predictable from the predictor variable. In multiple regression analysis, there can be P different predictor variables and C different criterion variables. Again, the square of the overall PC correlation indicates what percentage of the variability in the criterion variables is jointly predictable from the predictor variables. Each of the individual predictor variables also has a ("beta") weight that indicates what proportion of that overall predictability is contributed by that particular variable.

Now 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 mutiple regression equation of the form $b_1P_1 + b_2P_2 + ... b_pP_p = C$. 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. But the metrics will have been validated.

The UK Research Assessment Exercise

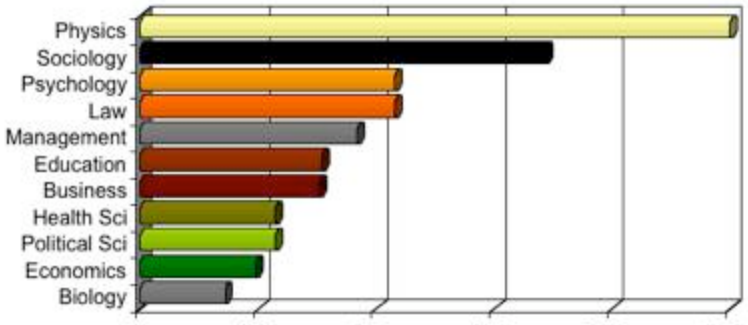
Is there any way this validation could actually be done? After all, journal peer review (as well as grant-proposal peer review) are done piece-wise, locally, and their referee ratings are both confidential and un-normalized. Hence they would not be jointly useable and comparable even if we had them available 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 United Kingdom's Research Assessment Exercise (RAE) (Harnad 2007; Butler 2008).

For over two decades now, the UK has assembled peer panels to evaluate and rank the research output of every active researcher in every department of every UK university every six years. (The departments were then accorded top-sliced research funding in proportion to their RAE ranks.) The process was very costly and time-consuming. Moreover, it was shown in a number of correlational studies that the peer rankings were highly correlated with citation metrics in all fields tested (Oppenheim 1996) – 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 2008 RAE 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. And as a bonus, generating and harvesting metrics on the Open Access Research Web will not only help measure and predict research performance and productivity: it 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 the articles that are made freely accessible to all would-be 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 Figure 1).



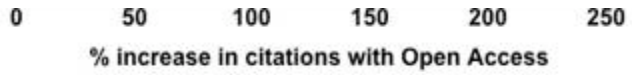


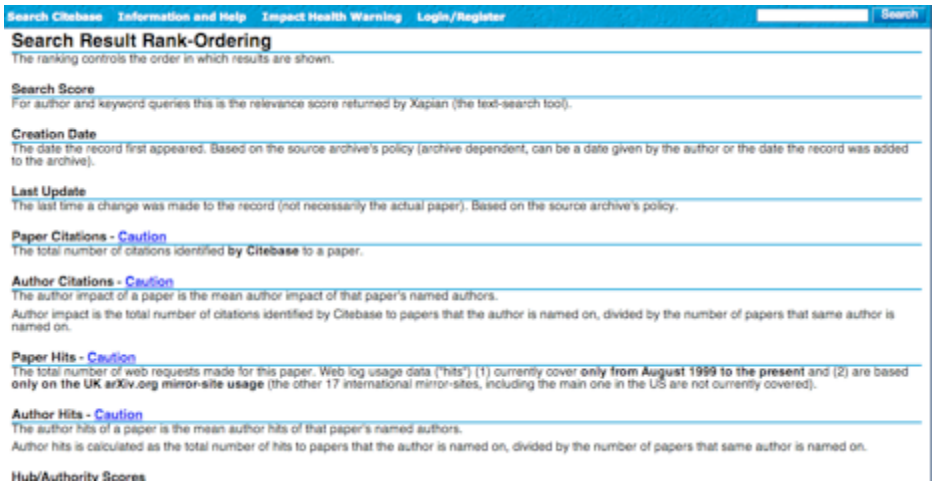
Figure 1. Percent increase in citations for articles (in the same issue and journal) that are made freely accessible online (Open Access, OA) compared to those that are not. The OA advantage has been found in all fields tested. (Data from Harnad & Brody 2004 and Hajjem et al 2005.)

There are many different factors contributing to this 'Open Access 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 downloading) and a *competitive advantage* (which will vanish once all articles are OA) – but 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 their usage and citation metrics (and their attendant rewards) is an incentive to researchers to provide Open Access 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.

Citebase : A Scientometric Search Engine

A foretaste of the Open Access Research Web is given by Citebase, a scientometric search engine (Brody et al 2006; Hitchcock et al 2003: http://www.citebase.org/). Based mostly on the Physics Arxiv, Citebase reference-links its nearly 500,000 papers and can rank search results on the basis of citation counts, download counts, and various other metrics (see Figure 2) that Citebase provides.


Search Citebase Information and Help Impact Health Warning Login/Register Search

Search Result Rank-Ordering

The ranking controls the order in which results are shown.

Search Score
For author and keyword queries this is the relevance score returned by Xapian (the text-search tool).

Creation Date
The date the record first appeared. Based on the source archive's policy (archive dependent, can be a date given by the author or the date the record was added to the archive).

Last Update
The last time a change was made to the record (not necessarily the actual paper). Based on the source archive's policy.

Paper Citations - Caution
The total number of citations identified **by Citebase** to a paper.

Author Citations - Caution
The author impact of a paper is the mean author impact of that paper's named authors.
Author impact is the total number of citations identified by Citebase to papers that the author is named on, divided by the number of papers that same author is named on.

Paper Hits - Caution
The total number of web requests made for this paper. Web log usage data ("hits") [1] currently cover **only from August 1999 to the present** and (2) are based **only on the UK arXiv.org mirror-site usage** (the other 17 international mirror-sites, including the main one in the US are not currently covered).

Author Hits - Caution
The author hits of a paper is the mean author hits of that paper's named authors.
Author hits is calculated as the total number of hits to papers that the author is named on, divided by the number of papers that same author is named on.

Hub/Authority Scores

Figure 2. Some of the metrics on which Citebase http://www.citebase.org/ can rank search results.

For a given paper, Citebase can also generate growth curves for downloads and the growth of citations (see Figure 3). It turns out that early download growth is a predictor of later citation growth (Brody et al. 2006).

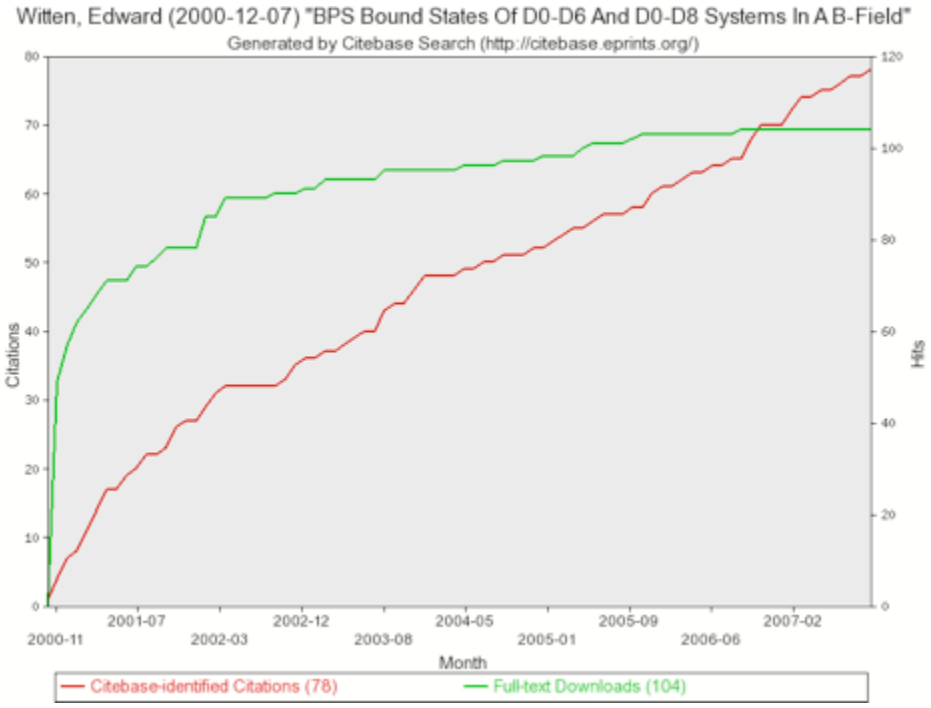


Figure 3. Citebase http://www.citebase.org/ growth curves for citations (red) and downloads (green) for a particularly important author in physics (E. Witten).

The various different metrics according to which Citebase can rank papers or authors can only be applied individually, one at a time in the current implementation. There is a menu (Figure 4: 'Rank matches byÉ') that allows the user to pick the metric. But in principle it is possible to redesign Citebase so as to rank according to multiple metrics at once, and even to adjust the weight on each metric. Imagining several of the vertical metric ranking options in Figure 2 arranged instead horizontally, with an adjustable weight (from -1 to +1) on each, gives an idea of how a search engine like this could be used to calibrate the outcomes of the multiple regression analysis described earlier for validating metrics. Exploratory analysis as well as fine-tuning adjustments could then be done by tweaking the beta

weights.

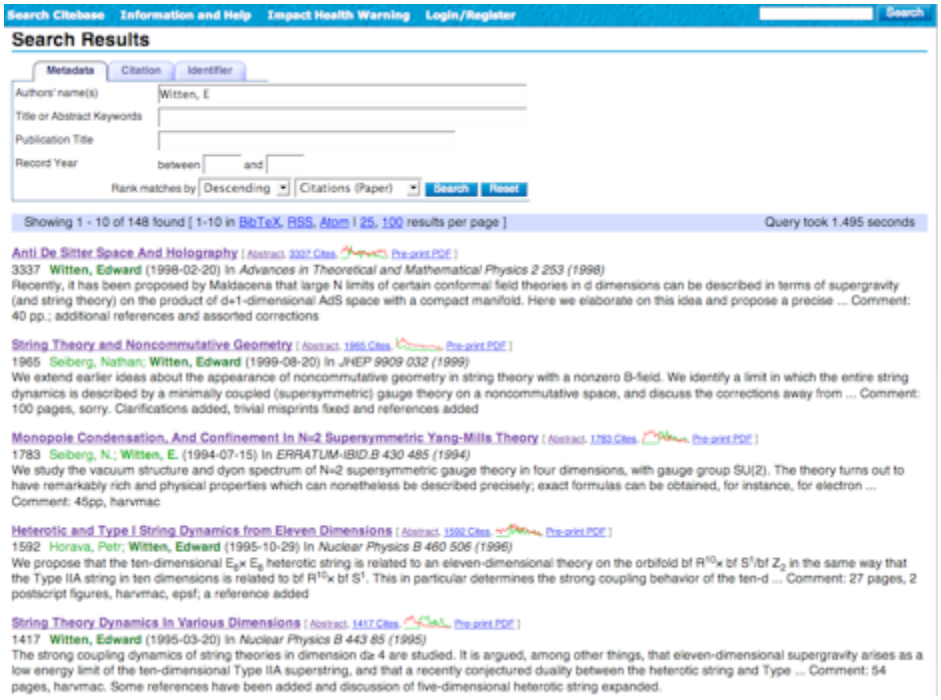


Figure 4. Citebase http://www.citebase.org/ allows users to choose the metrics on which they wish to rank papers, as well as to allowing them to navigate on the basis of of citation links.

LITERATURE CITED

Bradley, F.H. (1897/2002) *Appearance and Reality: A Metaphysical Essay.* Adament Media Corporation.

Brody, T., Carr, L., Gingras, Y., Hajjem, C., Harnad, S. and 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/

Brody, T., Harnad, S. and Carr, L. (2006) Earlier Web Usage Statistics as Predictors of Later Citation Impact. *Journal of the American Association for Information Science and Technology (JASIST)* 57(8) pp. 1060-1072. http://eprints.ecs.soton.ac.uk/10713/

Butler L (2008) Using a balanced approach to bibliometrics: quantitative performance measures in the Australian Research Quality Framework (this issue ESEP)

Campbell P (2008) Escape from the impact factor (this issue ESEP)

Hajjem, C., Harnad, S. and Gingras, Y. (2005) Ten-Year Cross-Disciplinary Comparison of the Growth of Open Access and How it Increases Research Citation Impact. *IEEE Data Engineering*

Bulletin 28(4) pp. 39-47. http://eprints.ecs.soton.ac.uk/11688/

Harnad, S (1985) Rational disagreement in peer review. *Science, Technology and Human Values.* 10 p.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. Pp. 235-242. http://cogprints.org/1646/

Harnad, S. (2004b) Enrich Impact Measures Through Open Access Analysis. *British Medical Journal* 2004; 329: http://bmj.bmjjournals.com/cgi/eletters/329/7471/0-h#80657

Harnad, S. & Brody, T. (2004) Comparing the Impact of Open Access (OA) vs. Non-OA Articles in the Same Journals, *D-Lib Magazine* 10 (6) http://eprints.ecs.soton.ac.uk/10207/

Harnad, S. (2007) Open Access Scientometrics and the UK Research Assessment Exercise. In Proceedings of 11th Annual Meeting of the International Society for Scientometrics and Informetrics 11(1), pp. 27-33, Madrid, Spain. Torres-Salinas, D. and Moed, H. F., Eds.
http://eprints.ecs.soton.ac.uk/13804/

Harzing AWK, van der Wal R (2008) Google Scholar as a new source for citation analysis (this issue ESEP)

Hirsch, Jorge 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

Hitchcock, Steve; Woukeu, Arouna; Brody, Tim; Carr, Les; Hall, Wendy and Harnad, Stevan. (2003) Evaluating Citebase, an open access Web-based citation-ranked search and impact discovery service
http://eprints.ecs.soton.ac.uk/8204/

Kleinberg, Jon, M. (1999) Hubs, Authorities, and Communities. ACM Computing Surveys 31(4)
http://www.cs.brown.edu/memex/ACM_HypertextTestbed/papers/10.html

Kline, Paul (2000) The New Psychometrics: Science, Psychology and Measurement. Routledge

Lawrence, S. (2001) Online or Invisible? *Nature* 411 (6837): 521
http://citeseer.ist.psu.edu/online-nature01/

Moed, H. F. (2005) Citation Analysis in Research Evaluation. NY Springer.

Oppenheim, Charles (1996) Do citations count? Citation indexing and the research assessment exercise, Serials, 9:155-61, 1996. http://uksg.metapress.com/index/5YCDB0M2K3XGAYA6.pdf

Shadbolt, N., Brody, T., Carr, L. and Harnad, S. (2006) The Open Research Web: A Preview of the Optimal and the Inevitable, in Jacobs, N., Eds. Open Access: Key Strategic, Technical and Economic Aspects, chapter 21. Chandos. http://eprints.ecs.soton.ac.uk/12453/

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

---

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"