Journal of Informetrics 10 (2016) 162–173

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

A comprehensive examination of the relation of three citation-based journal metrics to expert judgment of journal quality

Peter Haddawy[a,*], Saeed-Ul Hassan[b], Awais Asghar[b], Sarah Amin[b]

[a] Faculty of ICT, Mahidol University, 999 Phuttamonthon 4 Rd., Salaya 73170, Nakhonpathom, Thailand
[b] Information Technology University of the Punjab, 346-B, Ferozepur Road, Lahore 54700, Pakistan

ARTICLE INFO

Article history:
Received 12 August 2015
Received in revised form
15 December 2015
Accepted 15 December 2015

Keywords:
Citation analysis
Peer review
Source normalized impact per paper
Raw impact per paper
Journal impact factor
ERA

ABSTRACT

The academic and research policy communities have seen a long debate concerning the merits of peer review and quantitative citation-based metrics in evaluation of research. Some have called for replacing peer review with use of metrics for some evaluation purposes, while others have called for the use peer review informed by metrics. Whatever one's position, a key question is the extent to which peer review and quantitative metrics agree. In this paper we study the relation between the three journal metrics source normalized impact per paper (SNIP), raw impact per paper (RIP) and Journal Impact Factor (JIF) and human expert judgement. Using the journal rating system produced by the Excellence in Research for Australia (ERA) exercise, we examine the relationship over a set of more than 10,000 journals categorized into 27 subject areas. We analyze the relationship from the dimensions of correlation, distribution of the metrics over the rating tiers, and ROC analysis. Our results show that SNIP consistently has stronger agreement with the ERA rating, followed by RIP and then JIF along every dimension measured. The fact that SNIP has a stronger agreement than RIP demonstrates clearly that the increase in agreement is due to SNIP's database citation potential normalization factor. Our results suggest that SNIP may be a better choice than RIP or JIF in evaluation of journal quality in situations where agreement with expert judgment is an important consideration.

1. Introduction

Peer review has long been the standard for quality assurance in scholarly research. But because it is inherently subjective and qualitative, the reliability of peer review as a method to evaluate research quality has often been questioned. Issues include biases in selection of reviewers, the tendency of reviewers to evaluate according to their own interests, conflicts of interest, and biases in evaluating research (e.g. researcher age, university reputation) (Martin & Irvine, 1983; Smith, 1988; Langfeldt, 2001; Butler & McAllister, 2009). With the introduction of the Science Citation Index by Eugene Garfield in the 1960s academics, administrators, and research policy experts began to ask whether bibliometric indicators, primarily citation-based, might provide an alternative quantitative and more objective measure of research quality that would not suffer from the drawbacks of peer review (Garfield, 1979). This trend gained momentum with the introduction of the Scopus

* Corresponding author. Tel.: +66 2 441 0909.
E-mail address: peter.had@mahidol.ac.th (P. Haddawy).

http://dx.doi.org/10.1016/j.joi.2015.12.005
1751-1577/

and Google Scholar citation databases in 2004, as well as more sophisticated metrics and commercial analytical tools making use of the bibliometic data such as InCites (2015) and SciVal (2015). On one end of the debate, some academics argued that some uses of peer review, particularly for national research evaluation exercises, should be replaced with the use of bibliometric analysis. Arguments mainly focus on the enormous time and cost involved in the peer review process (Martin, 2011) and some strong correlations that have been shown between bibliometirc analyses and peer review (Butler & McAllister, 2009; Bertocchi, Gambardella, Jappelli, Nappi, & Peracchi, 2015). Abramo and D'Angelo (2011) have also argued that the labor intensive nature of peer review limits the number of research outputs that can be evaluated in large research evaluation exercises and that the limits on the sample size have negative effects in terms of robustness, validity, and functionality. But just as peer review has drawbacks and limitations, bibliometric measures have well known limitations as well. These include issues such as differences in types of research outputs among fields, variability in citation patterns across fields, the inability of citations to measure the quality of recent works, and the susceptibility of some indicators to manipulation. The realization that both peer review and bibliometric measures have different strengths and weaknesses has led some scholars to argue that pure peer review should be replaced by *informed* peer review in which reviewers are provided with quantitative bibliometric analyses. Arguments have focused on the complementary nature of the evaluation provided by quantitative indicators and peer assessment – Martin and Irvine's concept of "converging partial indicators" (Martin & Irvine, 1983) – and that metrics can provide tools to keep the peer-review process honest and transparent (Smith, 1988; Van Raan, 2003; Aksnes & Taxt, 2004; Moed, 2007). A consensus seems to have now formed around the idea of peer review informed by appropriate bibliometric analyses. As observed by Hicks, Wouters, Waltman, de Rijcke, and Rafols (2015), "Quantitative metrics can challenge bias tendencies in peer review and facilitate deliberation. This should strengthen peer review [. . .]. However, assessors must not be tempted to cede decision-making to the numbers."

But whether bibliometric indicators are used as an alternative to peer review or as supporting information for informed peer review, an important question is the extent to which they are, in fact, measuring the same thing. In other words, to what extent do bibliometric measures agree with peer judgments of research quality? In this paper we focus on the evaluation of journal quality and examine the extent to which the traditional Journal Impact Factor (JIF) and the newer Source Normalized Impact per Paper (SNIP) agree with expert judgement. While the JIF is still the most widely used bibliometric measure of journal quality, the SNIP addresses the important issue of differences in citation rates among fields (Moed, 2010; Waltman, Van Eck, Van Leeuwen, & Visser, 2013). While studies have been carried out comparing statistical properties of SNIP and JIF (Colledge et al., 2010), to date no extensive study has been carried out comparing the measures to human expert judgment. Additionally, in order to determine the role in any correlation played by SNIP's normalization for citation differences among fields, we include in our analysis the raw impact per paper (RIP), which is identical to SNIP without the normalization factor. For our study we make use of the Excellence in Research for Australia (ERA) journal rating which provides a database of over 20,000 journals spanning a broad array of fields, ranked by experts in the various fields. The ERA was an Australian national exercise to provide the government with an overview of strengths and weaknesses of universities in the country and the journal rating was introduced to help evaluate the quality of publications.

2. Related work

Vanclay (2011) presents an analysis of the ERA journal ranking, which includes a comparison with the original version of the SNIP indicator. While the work has been rather heavily criticized for various methodological flaws (Butler, 2011; Jasco, 2012), the similarity of the issues examined in that paper and our work warrants some discussion. Vanclay argues that the use of SNIP for comparison "offers some independence since it was published after public submission on the ERA ranking closed". The analysis focuses on the two fields of Design and Forestry. Vanclay shows that the correlation between the ERA rating and the SNIP values is 0.08 for Design and 0.52 for Forestry journals. While he does not provide a correlation analysis for the ERA list overall, he does show the log-average SNIP scores for each ERA tier for the 23 top level Field of Research (FOR) disciplines. In particular, he shows that for the discipline Built Environment and Design the log-average SNIP value for the A* tier is lower than for the A tier. He also shows great variation of the log-average SNIP values among the disciplines in the A* tier. His results are not precisely comparable with ours since his sample set contains only 9118 journals, about 2000 journals fewer than in our set.

Haddow and Genoni (2010) compare the ERA ranking of Australian Social Science journals with several citation-based indicators, including total citations, h-index, and impact factor. They examine only the Australian journals in this discipline and find little relation between the citation-based measures and the expert rating. They attribute this to a number of factors particular to Social Sciences, including long time lags before publications are cited, a significant proportion of publications appearing in non-journal outlets, and the importance of journals focused on national issues despite those journals attracting fewer citations.

Serenko and Dohan (2011) compare expert ranking of journal quality and impact in the field of Artificial Intelligence with citation based indicators. They conducted a survey and received 873 valid questionnaires from experts who were asked to rate each of 182 journals in terms of its overall contribution to the field of AI. They find moderate correlation between the citation-based impact measures and expert opinion. In particular, the correlation with 2009 JIF is 0.508. They find expert ranking results strongly influenced by the experts' major current research area even in this relatively narrow field. In addition, they compare the ERA rating of the set of journals with the 2009 JIF and find a correlation of 0.563.