Research Performance Evaluation Approaches and New Evaluation Development

Muzammil Tahira, Rose Alinda Alias and Aryati Bakri
Department of Information System, Faculty of Computing, Universiti Teknologi Malaysia, Skudai-31810, Johor, Malaysia

ABSTRACT
Quantity and quality dimensions of research performance using various indicators are measuring by applying different research performance evaluation approaches. Metrics/indices play an essential role for peer based, metrics based or hybrid approaches. The potential use of new research performance indices like h and h-type indices is the hottest topic among scientometricians and policy makers. A range of contradictory views and tensions have arisen regarding the use of metrics or peer review process and potentialities of the use of new indices. This review study is based on the similarities and contradictory concerns that are raised in policy level, empirical and point of views studies about RPE approaches, often used metrics and most noted development named h-index.

Keywords: Research Performance Evaluation, Metrics Based Approach, Peer Based Approach; Hybrid Approach, Publication, Citation, Journal Impact Factor and h-index.

I INTRODUCTION

Three relations are described between scientometric and peer review. These indicators can be used as a supplementary, complementary or as a validation instrument of peer review outcome (Moed, 2005). Research evaluation is a cumbersome work that requires objective and subjective aspects to handle. “Bibliometrics offer a powerful set of methods and measures to study the structure and process of scholarly communication” (Borgman and Furner 2002). In practice, bibliometric and scientometric are used as synonymous. In general research productivity dimensions are explored by quantity (number of publications) and quality (total citation counts) of output records with traditional activity and impact measures (Van Raan, 2004; Hirsch, 2007). Worldwide universities have a number of mechanisms for accountability and research evaluation. Ranking mechanism and indicators are ‘culture constructs’ (Leydesdorff, 2008). Different metrics are applied to research evaluation at various levels to measure the productivity (output) and impact. The Research Performance Evaluation (RPE) of United Kingdom, Hong Kong, Netherlands, Australia, Hungry, Poland, New Zealand (Geuna and Martin 2003) and Malaysia (Bakri, 2010) has a kind of peer assessment with ‘informed bibliometrics’ for decision making. The researchers are focusing on exploiting dataset, based on reference enhanced databases. The arbitrary nature of indicators, disciplinary perspectives, electronic publishing scenario and use of indicators in different contexts have turned the attention of this community, policymakers and researchers to discover new metrics, modification of existing one to gauge the quality and quantity challenges. The dynamics in the use of metrics is required to avoid from ‘homogenization’ in research output. They are examining the applications of traditional and new metrics in different contexts and aggregate level (Lazaridis, 2010; Imperial and Navarro, 2007; Moed, 2005; van Ran, 2005, Norris & Oppnehim, 2003; Oppnehim, 1997).

The present study will add to the growing body of knowledge. It will assemble a few salient aspects about the RPE approaches; often use tradition metrics (publications, citations and Journal Impact Factor) and helps in understanding the place of new metric, if any. It will end after raising many avenues of inquiry for the future study/research.

II OBJECTIVES AND METHODOLOGY

Based on review methodology, the study kindles to the RPE approaches, most used metrics and new developments. Firstly, different RPE approaches are discussed. First section describes the Peer Based Approach (PBA), Metrics Based Approach (MBA) and new perspective on hybrid approach. Next section describes the advantages and disadvantages of newly introduced and widely acceptable metric named h-index. The central stance is to seek the potential use of h-index for RPE purpose as discussed in literature.

III SECTION 1: TYPES, WEAKNESSES AND STRENGTHS OF RPE APPROACHES PEER BASED APPROACH (PBA)

Hackett (1997) has described four desirable outcomes expected from peer assessment. First and the most important are effectiveness (doing peer
review thoroughly) and efficiency (not taking too much time) and second is responsiveness (helping to identify new avenues of inquiry) and inertia (peer review is conservative because new ideas are judged against existing approaches). Other expectations are to examine the meritocracy (choosing the best) as well as fairness (for instance, for young researchers or women) and to the inclusions of halo effect (poor papers are easier for respected academics to publish). Peer review approach has been using globally and is widely acceptable. This is considered a direct and accurate help, to identify centre of excellence and the area of strength, to explore other factors, inclusion of international perspectives, international benchmarking, flexible and multifunctional approach (Butler and McAllister, 2011). This approach has a central position in RPE, and the role of scientometrics is to provide a useful support for decision making (Derrick, Haynes, Chapman and Hall, 2011; Butler, 2008; Moed, 2005; REPP, 2005).

PBA is mostly criticized due to use of ‘basket of measures’ (Butler and McAllister, 2011), more bureaucratic burden, discrimination against main stream search and difficult to find experts of all fields (Butler and McAllister, 2011; Donovan, 2007a; 2007b; Hirsch, 2005; Moed, 2005). Lack of validity and reliability issues in peer assessment are questioned due to subjectivity (Moed, 2005; Bakri, 2010; van Raan, 2005). This approach is considered more conservative (Hackett, 1997), lack in accountability and clear criteria (Lund University, n.d.). Three tensions were discussed by Donavan. One tension is ‘relevance gap’ another is “difficulty to distinguish between scholarly excellence and usefulness” pointed out by Nightingale and Alister Scott and, the third one is ‘polarization of the opinions into two camps (PBA and MBA)’ raised by Dovvan (2007b).

IV METRICS BASED APPROACH (MBA)

Various reports and authors have pointed out known flaws in MBA for instance; metrics do not actually measure research quality. ‘Research income’ is an input measure. Getting a contract and grant does not ensure high-quality outcomes (Donovan, 2005). This is often criticized due to document type (Moed, 2005, van Raan, 2005; Opthof and Leydesdorff,2009), modest incentives for cooperation or collaboration, some disciplinary area or institutions may get advantages (Donovan, 2005) as a rough proxy, encourage goal-displacing behaviour (DEST, 2005), trendy literature (Tahira, Alias & Bakri, 2011) and Science (STEM) vs. Social science (HASS) practices (Abramo and Angelo, 2011; Leydesdorff, 2008); van Raan, 2005 Moed, 2005; DEST, 2005; Donovan, 2007b).

“Industry grant, patents, intellectual property right items, do not actually measures the impact as they do not reflect the actual research outcome and benefits” Council for the Humanities, Arts and Social Sciences (CHESS, 2005).

Citations are found the most significant predictor of RPE (Butler and McAllister 2009). Research assessment based on citation analysis is not enough to deal with the excellence (Abramoand Anglo, 2011; Opthof& Leydesdorff, 2011; van Raan, 2005; Moed, 2005). Culture of citation varies from field to field, journal, books and other mode of publications. It does not reflect the inherent quality. Work may get more citation because its findings are contested and citing intention may be positive or negative (Donovan, 2005). High-impact journals are not a guarantee to get more citations (McRae-Spencer, 2007). The subject of the inclusion of self-citation or exclusion of the journal and authors is also a debatable issue in a different culture and disciplinary perspectives. The phenomenon described by Anthony van Raan referred as ‘sleeping beauty’ and ‘kissing princess’ plays an important role in citing and cited by (2004). A publication with few or zero citations in the previous years, and awake by any princess will start getting many citations in the near future. Evaluative bibliometric seeks citations as a measure of impact. Other limitations are ‘perfunctory’ incorrect work, errors in counting, time window, homograph problems, disciplinary perspectives, bias of database publishers towards language, regions USA vs. UK practices pointed out by Campbell et al. 2010).

Journal Impact Factor (JIF) has been established as a useful, influential and impact measure of journal. It is widely used as a proxy of a journal’s quality and scientific prestige (Bormann, Marx, Gasparyan, and Kitas 2011). Pendlebury (2009) described JIF as a simple measure that provides a global view as well as acceptance and put insight into the recent performance. Common criticism pointed out by him was inflation in the numerator, confusion and concern over the definition of citable items in denominator, lack of multidimensionality, article type, databases issue, matching of the journal name, citations practices, multidisciplinary subject issues, social science disciplines, bias against certain nations, language and two year time window. Though it becomes a “chief quantitative
measure of the quality of a journal, its research papers, the researcher(s), and even the institution” nevertheless, it cannot be used as a direct measure of quality (Amin and Mabe, 2003). They also reported many sociological and statistical factors that affect the IF such as the subject variations, journal type, average number of authors per paper, size of the journal, citation measurement time window etc. IF should not be confused with impact of an article or author (Garfield 1994).

Coleman mentioned evolving scholarly communication system, the open-access movement, and increasing globalization as pressing reasons for journal value rather than just impact. These reasons can also applicable to other metrics. It is failed to capture all the facets (Coleman 2007). Many studies reported that specialized journals had self-citations trend and issue of change in name (Garfield 1974; McVeigh, 2004). A study to examine the use of normalized journal impact indicator instead of Journal Citation Report (JCR by ISI) IF resulted in a more variation in the first indicator than the second one (Moed 2009). No correlation between a given journal’s IF and citations obtained by an article in that journal (Imperial and Navarro, 2007) is the key criticism. Subjectivity, validity and reliability of JIF are questioned as this is the indicator of journal, not an article.

MBA is regarded as objective (Butler and McAllister, 2011, Donovan, 2007a; 2007b; DEST, 2005; Hirsch, 2005), economical (Donovan, 2007b; Hirsch, 2005), more valid (Butler and McAllister, 2011; Moed, 2005; Bayer & Folger, 1966; CHASS 2005; Hirsch, 2005), reliable (Donovan, 2005; Moed, 2009), scientific (Leydesdorff, 2008; Donvon, 2005; DEST, 2005), time saving (Butler and McAllister, 2011, Abramo and Angelo, 2011; Campbell, 2005; DEST, 2005; Bayer and Folger, 1966), simple (Leydesdorff, 2008; Donovan, 2007a; Donovan, C. 2007b; Donovan, 2005; DEST, 2005), ease in use (Butler and McAllister, 2011, Abramo and Angelo, 2011; Campbell, 2005; DEST, 2005; Bayer and Folger, 1966) and persuasive (Butler & McAllister, 2011). While, PBA is subjective (Butler and McAllister, 2011; Donovan, 2005; REPP, 2005; Butler, 2003) and reliable (Butler and McAllister, 2011; Donovan, 2007b), but, it is weak incorporating other features.

Any metrics approach for performance evaluation should be used as a discipline-specific set of indicators (Donovan, 2005). Donovan regarded PBA “as an improper process, a black box, a process in isolation rather than the public interest” (Donovan, 2007a.), biased as a confounding factor (Pendlebury, 2009), subjective, reliant, messy and not replicable (as cited in Donovan, 2007b).

Lacking in objectivity, simplicity, requiring more time and expert panels, non-scientific approach and complication in processing and disciplinary perspectives are major shortcomings in MBA under eight constructs. It can be persuasive and become the cause of originating of new metric. MBA is often debated due to non-subjective nature, validity and reliability issues in the metrics. It is more persuasive in nature, and several variants of metrics have been introduced to add value and to complement. Standard metrics neither measure quality or impact nor are novel alternatives (Leydesdorff, 2008). Peer review should be supported by a variety of quantitative quality metrics (Moed, 2005; Braun, Glanzel and Schubert 2006). The future metrics would be more informed (Harnad, 2008), Butler and Moed visualized the future of RE as hybrid approach while, Donavon looked it a pure subjective in nature.

V  SECTION 2. THE H-INDEX

Legitimacy of often used metrics for RPE has been challenged (Lesdrohoff, 2009). Literature observed noteworthy fluctuation in ranking criteria with the application of new and traditional indicators (Schreiber, 2008). Such discrepancies in the results lead to evaluate the existing practice for the potential use of new development(s) and to explore for further improvement.

The major change in the scientometric research scenario is the proposed ‘h-index’ by Hirsch (2005). This index claims to measure the impact of scholarly communication in terms of quality (citation) and productivity (publications). The h-index indicates; ‘brute force in citations’ and the beauty of the new index lies in the blend of both cores. It deals with both quantity and quality issues in an objective manner, gains significant attention of the community and warmly welcomed. Since then, an ongoing debate by scientometricians, policy makers and researchers of the various disciplines made it a hot topic in the history of research.

Advantages and Disadvantages

A comprehensive view about h-index, its advantages and disadvantages, are presented in Table 1. This is based on empirical studies and viewpoints related to h-index and its application as a performance measure tool. These empirical
studies are conducted in different contexts, disciplinary perspectives and argument in favour and disfavour of the place of this indicator for RPE.

Numerous h-type indices introduced to overcome the pitfalls. Bornmann et al. (2008) indicated that indices other than the h index were even better suited to the said purposes. An overview of these h-type indices was well stated by Bornmann et al. (2012). They presented an overview of 37 h-type indices. These indices were computed to address the implicit and explicit shortcomings of original h-index. The study mainly deals with disciplinary perspective, age, co-authorship issues, robustness and citation distribution underpinnings.

In their meta-analysis study, they found a high correlation between various h-type indices in the range of 0.8 and 0.9 between the h index variants and a mean value. “Generally high correlation found between the h index and its variant does not imply that the relationship is valid for each individual case (P.11)”. In summary, the shortcomings of h-index are: ignoring highly cited publications and less below the h-index value, weak for immediate inclusion of publication, and a single number can not reflect all aspects of a good indicator, fail to incorporate overall performance of researchers, inconsistencies, same h-index issues and robustness are distinctive features of h-index that are criticized and taken as disadvantages. The field dependent nature, self citation, multi-authoredness, and career length are also taken as its demerits. Fairly these concerns are also related with other traditional or often used metrics.

VI DISCUSSION AND CONCLUSIONS

The research performance evaluation clearly emphasizes on peer review approach. In general, any change in the nature of RAE approach is due to technological, organizational and social changes, relation of the industry and academic, and dissatisfaction about previous systems or absence of systems. In the most countries, RE is still in nascent or experimental stage (Von Tunzelmann and Kraemer Mbula, 2003). Quantity, quality, efficiency, relevancy and visibility are major dimensions of university research evaluation (Campbell, 2005). Besides all limitations, anomalies and tensions, surprisingly, significant correlations have been observed between expert judgments and bibliometric data in the studies carried out in Italy (Reale, Barbara, Costantini, 2007; Abramo, D’Angelo, 2011, Caprasecca, 2009), the Netherlands (Rinia, Van Leeuwen, Van Vuren, and Van Raan, 1998; van Raan, 2005), UK (Norris and Oppenheim, 2003; Oppenheim, 1997; Seng and Willett, 1995), and weak correlation in Malaysia (Bakri, 2010).

Common approaches used for RPE are either metrics based or peer assessment. The mix mode approach is more multifunction and flexible. The term ‘impact’ evaluation for RPE is becoming a crucial issue due to e-publishing scenario, methodological issues and policies. Scientometric is an established monitoring tool for every approach. Metrics/indices performance is certainly not perfect but makes it better than peer review in terms of robustness, validity, functionality, costs and execution time.

Hybrid approach owns the strengths and weaknesses of both. This is more multifunctional and flexible but, certainly not as simple, economic, robust and becomes less scientific when compared to MBA. Inclusion of both subjective and objective dimensions makes it more reliable and valid. This leads to persuasiveness and a new avenue of inquiry. New trends are to be more informed and metric oriented peer assessment, economic and robust approach for RPE (Abramo & Angelo, 2011; Moed, 2008; Imperial & Navarro, 2007); CHESS, 2005). This requires more sophisticated metric to quantify objective and subjective aspects. The most popular and most noted development is h-index.

The beauty of the new index lies in the blend of both cores. It deals with both quantity and quality issues in objective manners. Ignoring the highly cited papers, can be a merit in case of trendy literature’, disciplinary perspective and specific document type. The most visible disadvantage is the account of co-authorship patterns. This has been now addressed by researchers Bornmann, Mutz, and Daniel (2007); Altmann, Abbasi, and Hwang, (2009) and Moed (2011). This index faces methodological concerns in the collection of data and precision like other metrics. Jacso wrote a series of articles and highlighted the issues faced on computing this index via Google Scholar, Scopus and Web of Science databases (2008a; 2008b; 2010).

The arbitrary nature of indicators, disciplinary perspectives, use of indicators in different contexts and electronic publishing scenario have turned the attention of this community, policymakers and researchers to discover new metrics, modification of existing one to gauge the quality and quantity challenges. There is a growing consensus on ‘small is not beautiful’, and there is no ‘one size fit for
Further, there is a sound buzz for the inclusion of this metric among the scientific community, database service providers and policy makers. Since h-index has made its place for Research Performance Evaluation (RPE); therefore, its disadvantages should also be taken into consideration. It could be argued that the possible potentialities of this index should be examined in different contexts as well as disciplinary and regional perspectives. Whether this elegant measure may be endorsed or may not for evaluation purpose with already used metrics is still a hot debate. Prolific authors (Norris and Oppenheim, 2010; Luz et al. 2008; Bouabid & Martin 2008; Imperial & Navarro, 2007; Mingers 2007; van Raan 2006) emphasized to carry on further studies in a different context to explore its theoretical, methodological and empirical lures for RPE.

Table 1. Advantages and Disadvantages of h-index.  

<table>
<thead>
<tr>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple easily understood and robust (Hirsch, 2005; Glanzel, 2006a; Vancaly, 2007; Lazaridis, 2010).</td>
<td>It rises, even after one stop publishing (Burrell, 2007; 2009), proportional to the length of a person’s career (Norris and Oppenheim, 2007; Burrell, 2007). Not proper to show the fall-off of production in later years (Sidiropoulos et al., 2007). Earlier career at disadvantage (Glanzel 2006b; Burrell, 2007).</td>
</tr>
<tr>
<td>Balance way of comparing (Braun, Glanzel and Schubert, 2006); provide broadly balanced and cumulative view of an individual impact (Norris and Oppenheim, 2010a).</td>
<td>Weakly sensitive for very highly cited paper (Egghe, 2006a; Norris and Oppenheim, 2010; Bornmann and Marx, 2011).</td>
</tr>
<tr>
<td>Gauge productivity, importance or impact, alternative to others (Hirsch, 2005; Glanzel, 2006a), predictive power (Hirsch, 2007).</td>
<td>A single number cannot reflect all the aspects (Van Raan, 2005). Lacks in necessary accuracy and precision (Lehmann et al., 2005).</td>
</tr>
<tr>
<td>A useful supplementary indicator and can be used for durable performance (Glanzel, 2006b).</td>
<td>Measures quantity and quality (Hirsch, 2005; Braun, Glanzel and Schubert, 2006; Costas and Bordons, 2007).</td>
</tr>
<tr>
<td>Measures quantity and quality (Hirsch, 2005; Braun, Glanzel and Schubert, 2006; Costas and Bordons, 2007).</td>
<td>Incorporates both quantity and visibility of publications (Egghe, 2006a; Egghe and Rousseau, 2006; Van Raan, 2006; Glanzel, 2006a; Bornmann et al., 2009).</td>
</tr>
<tr>
<td>Incorporates both quantity and visibility of publications (Egghe, 2006a; Egghe and Rousseau, 2006; Van Raan, 2006; Glanzel, 2006a; Bornmann et al., 2009).</td>
<td>Insensitive to zero</td>
</tr>
<tr>
<td>Consistently with regard to citations (Braun et al., 2006).</td>
<td>Convergent validity and valid for research performance (Hirsch, 2005a; Cronin and Meho, 2006; Bornmann and Daniel 2005b; Van Raan, 2006; Kelly and Jennions, 2006).</td>
</tr>
<tr>
<td>Balance way of comparing (Braun, Glanzel and Schubert, 2006); provide broadly balanced and cumulative view of an individual impact (Norris and Oppenheim, 2010a).</td>
<td>Reflect the publication patterns of discipline (Hirsch, 2005; Glanzel, 2006a).</td>
</tr>
<tr>
<td>Reflect the publication patterns of discipline (Hirsch, 2005; Glanzel, 2006a).</td>
<td>Potential application to the assessment of less output (Glanzel, 2006b).</td>
</tr>
<tr>
<td>Reflect the publication patterns of discipline (Hirsch, 2005; Glanzel, 2006a).</td>
<td>Formal metric describe broad impact in an unbiased way (Hirsch, 2005). It can apply at all level of aggregation (Hirsch, 2005; Glanzel, 2006a; Schubert, 2007) and going to become widely acceptable and formal use (Bornmann and Daniel, 2007; Ball, 2007; Bornmann et al., 2011).</td>
</tr>
</tbody>
</table>

**ACKNOWLEDGMENT**

Authors would like to acknowledge the partial support in the form of International Doctoral Fellowship (IDF) from MOHE and UTM, Skudai, Johor, Malaysia and Higher Education Commission (HEC) of Pakistan.

**REFERENCES**


Lazaridis T. (2010). Ranking university departments using the mean h-index. Scientometrics, 82, 211.


The Times Higher Education assessed June 8, 2012 from www.timeshighereducation.co.uk.


