Library & Information Science Research 37 (2015) 51–60

Contents lists available at ScienceDirect

Library & Information Science Research

# Average evaluation intensity: A quality-oriented indicator for the evaluation of research performance[☆]

Zhiqiang Wu

*Department of Management Science, Sun Yat-sen Business School, Sun Yat-sen University, Guangzhou, 510275, PR China*
*Research School of Accounting & Business Information Systems, College of Business and Economics, Australian National University, ACT 0200, Australia*

ARTICLE INFO

Available online 7 February 2015

*Keywords:*
Research performance evaluation
Citation analysis
Citation motivation
Research quality
Scientific indicators

ABSTRACT

A variety of indicators have been created to measure the research performance of journals, scientists, and institutions. There has been a long-running debate on the use of indicators based on citation counts to measure research quality. The key argument is that using indicators based on raw citation counts to evaluate research quality lacks measurement validity. Traditional reference formats do not present any quality related evaluations of the citing authors toward their references. It can be argued that the strength of peer evaluation to a research output, which is taken to represent its quality, is the elementary unit in the evaluation and comparison of research performance. A good candidate for evaluating a piece of research is a researcher who cites the research and knows it well. By accumulating different citing authors' evaluations of their references based on a uniform evaluation scheme and synthesizing the evaluations into a single indicator, the qualities of research works, scientists, journals, research groups, and institutions in different disciplines can be assessed and compared. A method consisting of three components is proposed: a reference evaluation scheme, a new reference format, and a new indicator, called the average evaluation intensity. This method combines the advantages of citation count analysis, citation motivation analysis, and peer review, and may help to advance the debate. The potential advantages of and main concerns about the proposed method are discussed. The proposed method may serve as a preliminary theoretical framework that can inspire and advance a quality-oriented approach to the evaluation of research performance. At the current stage, it is best to treat the proposed method as speculation and inspiration rather than as a blueprint for practical implementation.

## 1. Introduction

Greater competition for limited academic resources results in more selective decision-making processes in relation to research resource allocation (King, 1987). Ideally, the award of honors, appointments, promotions, and funding allocation, all of which involve millions of researchers and billions of dollars in research funding every year, is based on candidates' scholarly merit. However, it is a difficult task to evaluate the performance of scholarly merit of research works, scientists, and research entities (Bridges, 2009). On the other hand, the criteria used in research performance evaluations influence how scientists conduct research, which further influences the progress of scientific research (Alberts, 2013; Fanelli, 2012; Moed, 2007). Thus, the quality of the methodology used in research performance evaluation has an impact on the progress of science (Nallamothu & Lüscher, 2012).

## 2. Problem statement

The current methods used in evaluating the performance of research works, scientists, journals, research groups, and institutions have prominent weaknesses. Great progress has been made in attempting to assess research impact using indicators based on citation counts. However, such measures have not captured the essence of research quality. Using indicators based on raw citation counts to evaluate research quality lacks measurement validity. A main criticism is that these indicators (and other methods with the exception of peer evaluation) basically look at the correlation between extrinsic features with research quality but do not explore intrinsic characteristics of research quality (Bridges, 2009). Combining multiple indicators does not improve this situation and hence cannot substantially enhance research evaluation. The use of these measures in the evaluation of research performance may drive researchers to change the extrinsic features of their work but do nothing to improve research quality (Bridges, 2009). To date, there is no single measure or approach that can evaluate research performance effectively.

Although the evaluation of research performance concerns wide interests and has been broadly explored, little research has directly and systematically addressed the basic questions: what should be measured

[☆] This research was conducted at Sun Yat-sen University and the Australian National University.

*E-mail address:* zsujimmy@163.com.

http://dx.doi.org/10.1016/j.lisr.2013.10.009
0740-8188/© 2015 Elsevier Inc. All rights reserved.

to evaluate research performance and how should it be measured? The present research starts from these basic questions to reconsider the issue of research performance evaluation. It explores the possibility of developing a single measure that can evaluate research performance effectively and efficiently. Following the line of research on using citation analysis to assess the performance of research works, scientists, journals, research groups, and institutions, this research suggests a potentially more effective approach to resolving this problem that has been debated for decades. This research will help people gain deeper understanding of the issues related to research performance evaluation by redirecting attention away from the current use of impact-oriented indicators toward quality-oriented indicators, and suggesting that the process of research performance evaluation should be more interactive, transparent, and public.

## 3. Literature review

Taking research activity as a production procedure, the methods used in the evaluation of research performance have evolved in four main categories: input-based, output-based, usage-based, and judgment-based (Table 1). First, research needs inputs of human labor, funds, and facilities. The indicators based on research inputs, such as calculating the sum of research funds and number of active researchers (Martin & Irvine, 1983), represent a resource-based view of research performance. This approach assumes that research resources are allocated based on the scholarly merit of research work, scientists, and/or research entities. It is expected that better research is more likely to get support. The more research resources an entity controls, the stronger its research ability. On the other hand, the indicators based on research outputs take a productivity-based view of research performance. They consider that the number of research outputs an entity can produce is more important than the research resources it has. They assume that publication is a good outlet for presenting research results and a good way to exhibit scholarly merit. The more publications, the greater the contribution made to the progress of science and society. For example, a common method is to calculate the number of publications (per researcher) (e.g., Martin & Irvine, 1983), especially within a certain spectrum of recognized academic journals (so-called top or leading journals) (e.g., Clark, Au, Walz, & Warren, 2011). Such indicators assume that all publications, especially in the same outlets, have equal scholarly merit. But these assumptions are not true (Martin & Irvine, 1983). Top research can be found in low-ranking journals (Hussain, 2011), while articles appearing in top journals may have problems or even serious mistakes. The perceived quality of the papers and the rankings of the journals in which they appear are not consistently correlated (Paul, 2008).

Research does not end with the publication of scientific reports or articles. In a broad sense, the research cycle should include readers' feedback, which may prolong the final judgment on the merit of a piece of work. The indicators based on research outputs do not show how readers respond to a researcher's works. A number of indicators have been proposed based on people's usage of research outputs, as reflected by their behavior, such as downloads (Bollen & Sompel, 2008) and readings (Darmoni, Roussel, Benichou, Thirion, & Pinhas, 2002). The indicators based on usage data measure some aspects of research impact on society, and are assumed to be associated with scholarly merit. In addition to those peripheral usages, an important usage of research outputs is citation, which represents a major form of usage of research outputs in the scholarly community.

Studies of the relative importance of scientific journals based on reference counts originated more than 80 years ago and initially aimed to help college librarians in the United States select journals for their collections (Archambault & Larivière, 2009; Gross & Gross, 1927). Following this inspiration, in around the 1950s to 1970s, the journal impact factor (average number of citations per item published within a specific period of time, e.g., two years) was developed to evaluate the importance of scientific journals (Garfield, 1955, 1972). Since then, other

**Table 1**
Major types of methods used in research performance evaluation.

| Category | General assumption | Typical indicator/method | Main assumption | Example | Applicable measurement level/unit | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | An | Wr | Rs | Gr | Jr | In | Nt |
| Research input-based (resource-based view) | Research resources are allocated based on scholarly merit. The more research resources one has, the stronger one's research ability. | Research fund/facility count-based | Research funds/facilities are allocated based on scholarly merit. | Garcia & Sanz-Menéndez (2005) | | | ● | ● | | ● | ● |
| | | Researcher count-based | Each researcher (with the same title/honor) has equal scholarly merit. | Pouris (1989) | | | | ● | | ● | ● |
| Research output-based (productivity-based view) | Publication is a good outlet for presenting research results and a good way to exhibit scholarly merit. | Publication count-based | Each publication has equal scholarly merit. | Martin & Irvine (1983) | | | ● | ● | | ● | ● |
| | | Publication/publisher reputation-based | The scholarly merits of publications and the ratings of their outlets match each other. Publications in the same outlets or outlets with equal reputations have equal scholarly merit. | Clark et al. (2011) | | ● | ● | ● | ● | ● | ● |
| Behavioral usage-based (impact- oriented) | Scholarly merit is associated with research impact, which can be reflected by users' behavioral records of usage of research outputs. | Readership count-based | Publications are read based on their merit. Each reader treats a publication positively and equally. | Darmoni et al. (2002) | | ● | ● | ● | ● | ● | ● |
| | | Citation count-based | Publications are cited because of their scholarly merit. Each citation treats a cited work positively and equally. | Garfield (1972) | | ● | ● | ● | ● | ● | ● |
| Cognitive judgment-based (quality-oriented) | Research quality, the degree of excellence of research or scholarly merit, can be reflected by peers' cognitive judgment of goodness on research outputs. | Traditional peer review | Research quality can be sufficiently identified by a (small) group of selected experts within a (limited) period of time. | Martin & Irvine (1983) | ● | ● | ● | ● | ● | ● | ● |
| | | Open peer evaluation | Research quality is better assessed in an open, longitudinal way. | The present research | ● | ● | ● | ● | ● | ● | ● |

1. Abbreviations for measurement levels/units from low to high aggregations: An = analytical; Wr = individual research work; Rs = individual researcher; Gr = research group; Jr = publication outlet/journal; In = institution; Nt = nation/region.
2. The applicable measurement levels mean that the addressed indicators have been or may be used to measure research performance at these levels. It does not mean that the author of the present research supports all these practices. For example, it is questionable to use citation count-based indicators at low aggregation levels.
3. The basic characteristics of the above methods/indicators can be combined to form composite methods/indicators, e.g., numbers of highly cited publications (Martin & Irvine, 1983).

quantitative indicators based on citation counts have been proposed (Pendlebury, 2009). In particular, the Eigenfactor and h-index have been used to assess academic journals' and scientists' performance (Bergstrom, West, & Wiseman, 2008; Bornmann & Daniel, 2008; Hirsch, 2005). Indicators based on citation counts have been widely used to evaluate the impact of articles, scientists, journals, research groups, universities and their departments, and nations (Borgman & Furner, 2002; Moed, 2005), as well as their quality in general (Warner, 2000).

The basic assumptions supporting the use of indicators based on citation counts to measure research quality are that authors cite previous studies because of their scholarly merits (Nieminen, Carpenter, Rucker, & Schumacher, 2006) and they treat each cited work positively and equally. However, citations may not represent scholarly merit, as many factors affect an author's citing behavior (Brooks, 1986; Camacho-Miñano & Núñez-Nickel, 2009), and some have no apparent relationship with quality, such as the accessibility of the journals and the language of the articles (Borgman & Furner, 2002; Bridges, 2009; Moed, 2005). Indicators based on citation counts have an inherent weakness, namely that they disregard citation motivations. The results of a number of previous studies on citation functions and motivations do not support the basic assumptions of indicators based on citation counts. For example, Teufel, Siddharthan, and Tidhar (2006a) proposed an annotation scheme for the citation function and found that more than 60% of citations were neutral. Kacmar and Whitfield (2000) examined how each article that was referenced was used by the citing authors in two important journals in management science. They found that the vast majority of the articles just listed the references, rather than using them as an integral component. Such references should not be regarded as being as important as those which serve a specific function in the development of an author's argument. The ranking of papers based on the number of central citations (e.g., testing ideas posited in the cited article) was different from the ranking based on the total number of citations. Thus, it is “impractical to use the citation as a unit of measurement for the study of information transfer and information use” (Cano, 1989, p. 289). Furthermore, un-cited work is inappropriately considered to be unnecessary, unused, or to have less merit than cited work (MacRoberts & MacRoberts, 2010).

Many scientists and editors have questioned the rationality of using indicators based on citation counts to measure research quality (Frey & Rost, 2010; Kostoff, 1998; Leydesdorff, 2008; Seglen, 1997; Smith, 1981) and deem them arbitrary (Colquhoun, 2003; Editorial, 2006). Some empirical studies found no correlation between an article's research quality as judged by expert peers and the number of citations received by that article (Nieminen et al., 2006; West & Mcllwaine, 2002). Highly cited papers do not necessarily report breakthrough research or even make a significant contribution (Aksnes & Rip, 2009; Tijssen, Visser, & van Leeuwen, 2002). Although various indicators based on citation counts have been suggested, mathematical improvement alone cannot increase the validity of such indicators (Leydesdorff, 2008) if the original variables do not capture the essence of research quality. Rather, they measure only a fraction of the research influence in society (MacRoberts & MacRoberts, 2010; Nallamothu & Lüscher, 2012).

“Impact may be easy to measure and audit, but relevance (i.e., whether the research meets societal requirements) is not” (Nightingale & Scott, 2007, p. 547). In view of the coarseness of indicators based on citation counts, some scientists argue for a return to using peer review in evaluating research performance (Alberts, 2013; Ball, 2005; Editorial, 2005; Lehmann, Jackson, & Lautrup, 2006; Willcocks, Whitley, & Avgerou, 2008). Traditional peer review processes are usually undertaken by a (relatively small) group of selected experts within a (relatively short) period of time. Peer review plays a key role in determining social relevance and scientific excellence and is the main form of decision making related to the evaluation of research performance (Nightingale & Scott, 2007). Traditional peer review is an evaluative approach based on human cognitive judgment. It assumes that scholarly merit can be adequately identified by such a small-scale cognitive appraisal process. However, traditional peer review also has significant weaknesses. It is qualitative, high cost, time-consuming, subjective, and not transparent. Particularly when there are many research works that need to be assessed, peer review is ineffective (Pendlebury, 2009). In addition, peer review does not always do a good job on emerging topics, when handling a broad range of views, or identifying genuinely excellent research (Moed, 2007; Nightingale & Scott, 2007; Zitt & Bassecoulard, 2008). The expertise and representativeness of peer reviewers have been questioned (Bence & Oppenheim, 2004). It is difficult to identify and appoint the most appropriate reviewers who are familiar with the research subjects being measured and who can give fair judgments (Abramo & D'Angelo, 2011). Moreover, peer judgment may be influenced by social, political, and cognitive factors other than scholarly merit (Moed, 2005). Thus, some researchers may consider traditional peer review as “no more than a partial indicator of contributions to scientific progress” (Martin, 1996, p. 350). In view of this dilemma, a number of researchers support the use of bibliometric indicators as complements or supplements to traditional peer review in the research assessment process, to compensate for its limitations (Moed, 2007; Pendlebury, 2009; van Raan, 2005b), reduce the workload of peer review panels, and mitigate the problem of implicit bias (Taylor, 2011).

## 4. A quality-oriented approach

### 4.1. Five basic questions on research performance evaluation

To address the issue of research performance evaluation, five basic questions should be considered. For all science performance evaluation indicators, there is a fundamental question: what is being measured and compared? The primary goal of conducting research is to produce new scientific knowledge that can advance scientific progress (Martin & Irvine, 1983). Scientific knowledge includes basic knowledge (improving human understanding of real world phenomena) and applicable knowledge (useful and satisfies societal needs) (Frey & Rost, 2010; Lee, 2006; van Raan, 2005b). A piece of research should address what scientific knowledge it produces and how it produces that scientific knowledge. Therefore, the quality of a piece of research can be understood as its substantive contributions to scientific knowledge, enhancing people's understanding of the world and/or solving actual problems based on rigorous, original, and error-free work, presented in a clear, logical manner. The most important concern in the evaluation of research performance is how many contributions a research work, scientist, or an entity can make or has made to scientific knowledge. Research quality is also called scholarly merit (Andres & Wang, 2012) and “should be considered the essence of scientific research” (Frey & Rost, 2010, p. 2). Based on this definition of research quality, it is generally believed that higher quality research is more likely to contribute effectively to desired social outcomes than lower quality research (Moed, 2007). In contrast, research impact generally refers to “a recorded or otherwise auditable occasion of influence from academic research on another actor or organization” (Public Policy Group, 2011, p. 11). It is extrinsic and influenced by the quality of a piece of research. Thus, research quality is more essential than research impact in research evaluation. Therefore, as research outputs are the products and outcomes of research activities, it is the quality of research outputs, not their research impact, which should be measured and compared in the evaluation of the performance of research and its components.

Quality has twofold nature. One component is objective, and is the inherent quality of the work. The other component is subjective, which means that it depends on the recognition or judgment of other scientists (Eppler, 2006; Martin & Irvine, 1983). “Quality assessments require human judgment” (Pendlebury, 2009, p. 6). What constitutes scientific knowledge depends on the cognitive evaluation (Moed, 2007) and recognition of the scientific community (Lee, 2006). Thus, it is appropriate to assume that the measure of the judgments of other scientists is a valid reflection of research quality. More specifically, peer recognition of research, which is taken to represent its research quality, should be measured and compared. “In order to make a fair comparison one should first define a cognitive unit of analysis” (Leydesdorff, 2005, p. 1511). In particular, the evaluation intensity of peer recognition of research works is an appropriate cognitive unit for the evaluation and comparison of research performance.

A subsidiary question is: who should evaluate research quality? Whether the knowledge is good and satisfies needs is subject to the judgment of those who know about and use it. The people who are eligible to evaluate a piece of research conducted by a researcher are those who have examined the research or used the research findings, and are knowledgeable about the field of the research, as well as the debates in the field and other background to the research. A good candidate for such a task is researchers who cite this research. It is true that not all citing authors know the research that they cite well. Therefore, only in the situation where citing authors know the work well should they be eligible to evaluate its research quality.

The third question is: how should research quality be evaluated? To ensure that the evaluations made by researchers in different disciplines are intelligible and comparable, a consensus on the format of research quality evaluation to be used by eligible citing authors would need to be reached. Research quality has multiple aspects and can be evaluated using a single dimension or using multiple dimensions. The present research adopts the single dimension approach, assuming that the quality of a piece of research can be evaluated by a single, synthesized dimension, further described below.

The fourth question is how to provide evaluation information. Citing authors can indicate evaluations of the cited works along with the references. To accommodate evaluation information, a new reference format should be adopted. Alternatively, citing authors could submit the evaluation information of the cited works along with the references to a third party processor (e.g., an authoritative editor or website). In the latter case, the evaluations of the cited works would not appear in the citing authors' works. The present research takes the first approach.

The fifth question is how to calculate the indicator. Accumulating all the judgments given by different citing authors toward a reference would produce a summary evaluation of the cited work. A common way to process the accumulated evaluation is to calculate the average evaluation intensity (AEI). The AEI can be used to evaluate research works, scientists, journals, research groups, and institutions.

Taking these questions into account, the proposed method has three components: a reference evaluation scheme, a reference format, and the AEI indicator.

### 4.2. Reference evaluation scheme

All citations have a linking function, while not all of them have an evaluative function. Citations can be classified into two classes: evaluative citations, which have an evaluative function, and non-evaluative citations, which only act as linkages between citing and cited works. This classification is consistent with that of Borgman and Furner (2002) who categorized link analysis into evaluative link analysis and relational link analysis.

Those citations where citing authors give a positive, negative, or neutral judgment about references have an evaluative function. Positive judgments mean that citing authors agree, get similar results, or reinforce the cited works' propositions. Negative judgments mean citing authors find errors or shortcomings in the cited works, or disagree with or criticize the cited works' propositions. Neutral judgments are also evaluations of cited works. Neutral opinions reflect that citing authors do not show either positive or negative attitudes toward cited works or that the total strengths of positive and negative attitudes balance, which can be considered as a zero evaluation. Non-evaluative citations mean that citing authors have no specific attitudes toward the cited works, for example using them just as information, or for comparison, or abstaining from evaluation.

Although the motivation for citing a work is complicated and various, generally speaking, the attitudes of citing authors toward their references can be generalized and classified into three categories: positive, negative, and neutral. The three categories are consistent with the three-group classification of citer motivations suggested by Brooks (1986) and the conflated citation function scheme of Teufel et al. (2006a). In addition, the attitudes of the authors toward a cited work may have different intensities, such as strong, medium, or weak. Therefore, including non-evaluative judgments as a category, it is possible to formulate a four-category, three-scale research evaluation scheme (RES) to be used by citing authors to evaluate their references. The 4 × 3 model is used to illustrate the proposed method (Table 2).

The citing authors use a code from the RES to express an evaluation toward a reference. A code contains a letter (or character) and a number. For example, a code of P3 (or $+3$) means the citing authors strongly support the research reported in the cited work, while N1 (or $-1$) means they slightly reject it. In the case of a neutral judgment, the code is “Z”. A code of “Z” represents an evaluation of zero or a situation where positive and negative judgments balance. A code of “A” or no code represents non-evaluation or abstention from evaluation. The codes, with the exception of “A”, are used in the calculation of average intensity.

### 4.3. Annotation formats in RES

Citing authors give an overall evaluative judgment to each of the references well known to them. These references will have a code of judgment. The citing authors add a code indicating the evaluative judgment behind the reference in the reference section. For example:

> *Lawrence, S., Pennock, D. M., Flake, G. W., Krovetz, R., Coetzee, F. M., Glover, E., Nielsen, F. A., Kruger, A., & Giles, C. L. (2001). Persistence of Web references in scientific research. Computer, 34(2), 26–31.* **[P2]**

To make the evaluation more understandable, the citing authors can provide a brief explication for their judgment behind the evaluation code. Besides the general evaluation approach, the citing authors can also give a code indicating judgment each time a reference is quoted, indicating specific evaluation of the quotation, e.g., a statement or an element in the cited work. That means a reference may have more than one code if it has been quoted more than once. Currently, there are two types of reference styles used in most publications: numbered references and alphabetical references. An example in the alphabetical reference style would be as follows:

> *Most of the invalid URLs could be relocated given more time and search tools (Lawrence et al., 2001, P2; Casserly & Bird, 2003, P2). All of these findings are confirmed in the present study.*

In the numbered reference style it would appear as follows:

> *Most of the invalid URLs could be relocated given more time and search tools*[3, P2; 8, P2]*. All of these findings are confirmed in the present study.*

**Table 2**
Reference evaluation scheme.

| Category | Description | Abbreviation | Intensity | | |
|---|---|---|---|---|---|
| | | | Weak | Medium | Strong |
| | | | 1 | 2 | 3 |
| Positive | Agree, reinforce | P | P1 | P2 | P3 |
| Negative | Disagree, criticize | N | N1 | N2 | N3 |
| Neutral | Zero evaluation | Z | Z | | |
| Non-evaluative | Abstain, inform, contrast | A | A or without a code | | |

An evaluation code representing a citing author's overall evaluation of a cited work is applicable for the analysis at the level of an individual work or above, such as the evaluations of research works, scientists, or journals. An evaluation code representing a citing author's evaluation of a quotation from a cited work is used for the analysis at an analytical level. As, in most cases, people evaluate and compare research performance at or above the level of an individual work, the present research focuses on the use of evaluation codes representing citing authors' overall judgments of cited works to generate evaluation indicators.

### 4.4. Evaluation indicators

The AEI can be defined and then used to evaluate the research quality or performance of individual research works, scientists, journals, research groups, and institutions, etc. The values of AEI, called $W_{AEI}$, meaning the weight of AEI, are based on the values of cumulative evaluation intensity (CEI). The values of CEI and AEI vary according to different situations.

In most cases of research performance evaluation, the basic unit of analysis is the research work. To assess the quality of an individual work, all citing authors' evaluations of this work are aggregated. These evaluations can be represented by a table (Table 3). The variables $a_1$, $a_2$, $a_3$, $b_1$, $b_2$, $b_3$, and c represent the numbers of evaluation codes.

The value of CEI ($V_{CEI}$) of a reference as a summary of all citing authors' evaluations is the difference between the total intensities of positive judgments and the total intensities of negative judgments within a time period. The number of evaluation codes (or evaluative citations as one citation gives a general evaluation code to a reference) an individual work has received ($N_e$) can be calculated. The value of a reference's AEI per evaluative citation, called $W_a$, is the result of $V_{CEI}$ divided by $N_e$. $W_a$ is the average evaluation intensity given by citing authors to the same reference. If $W_a$ is greater than zero, the overall evaluation is positive. Otherwise, if $W_a$ is less than zero, the overall evaluation is negative. The more $W_a$ a work has, the higher quality it is. If no evaluative citation has been received, then the value of $W_a$ is defined as unavailable, and this will not affect the overall score.

$$W_a = \frac{V_{CEI}}{N_e} = \frac{\sum_{i=1}^{3}(a_i - b_i) \times i}{\sum_{i=1}^{3}(a_i + b_i) + c}$$

To evaluate the research performance of individual scientists, the following method would be used. Suppose a scientist has k works in a specific area within a time period. The values of $W_a$ of each of those works can be calculated according to the above method. Then the average evaluation intensity per work, called $W_s$, can be calculated. The $W_s$ of a scientist can reflect the average quality of research works the scientist has published. The higher value of $W_s$ a scientist has, the higher the quality of the scientist's research.

$$W_s = \frac{\sum_{i=1}^{k} W_{ai}}{k}$$

**Table 3**
Aggregated evaluations of a research work given by citing authors.

| Evaluations | Intensity | | |
|---|---|---|---|
| | Weak | Medium | Strong |
| Positive | $a_1$P1 | $a_2$P2 | $a_3$P3 |
| Negative | $b_1$N1 | $b_2$N2 | $b_3$N3 |
| Neutral | cZ | | |

To assess an academic journal a similar method can be used. Suppose there are t (research) articles that have been published in the journal within a time period, e.g., three years. The value of each article's $W_a$ is obtained according to the above method. Then the average evaluation intensity per article, called $W_j$, can be calculated and compared with those values in other journals. The $W_j$ of a journal can represent the average quality of (research) articles it has published. The higher the value of $W_j$ a journal has, the higher the quality it has.

$$W_j = \frac{\sum_{i=1}^{t} W_{ai}}{t}$$

The research performance of research groups, departments, institutions, or other organizational entities can also be assessed by this means. Suppose there are r members in a group. The value of each member's $W_s$ can be calculated according to the above method. Then the average evaluation intensity per person, called $W_p$, can be calculated and compared with the evaluation intensity of researchers in other groups. The $W_p$ of a group or department or an institution can represent the average quality of research works each member has published. The higher the value of $W_p$ a group has, the higher the quality of the group's research

$$W_p = \frac{\sum_{i=1}^{r} W_{si}}{r}$$

The values of all $W_{AEI}$ range from a maximum value to a minimum value according to the evaluation scale used. In the case of three scales, the ranges of $W_{AEI}$ are between $+3$ and $-3$. It should be noted that in calculating the above indicators, those research works not receiving evaluative citations and their authors should be omitted.

Changes in the value of $W_{AEI}$ over time can also be observed, and can be used to assess changes in research performance. For instance, even though some overall $W_{AEI}$ may be unsatisfactory, if the curves of $W_{AEI}$ (e.g., $W_a$ for individual works) show upward movement over time (Fig. 1), it can be inferred that the propositions advanced in a paper were not accepted when they first appeared, but gained recognition afterwards. In contrast with this, downward curves in $W_{AEI}$ over time may indicate that some flaws were found over time. Usually, new citation evaluations are more valuable than old ones. Thus, the latest part of the curve will be more valuable.

In addition, if the values of multiple $W_{AEI}$ being compared are the same, some tactics can be used to differentiate them. For example, evaluators can compare their variances or standard deviations, which can also be calculated. Variance and standard deviation are often used to measure the average scatter around the mean in statistical data (Lee & Lee, 2009). A smaller variance or standard deviation indicates that the evaluations given by citing authors are more convergent. Otherwise, evaluators can recalculate and compare their $W_{AEI}$ over a shorter, but more recent period, because newly received citation evaluations are more valuable in general. Furthermore, it is also possible to calculate

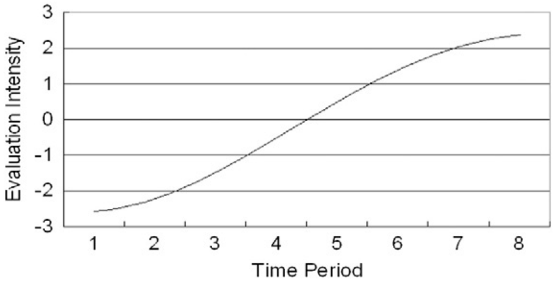

**Fig. 1.** A simulation of an upward curve of $W_{AEI}$.

the percentile ranks of $W_{AEI}$ within an area or a scope and use them as supplementary indicators.

## 5. Discussion

### 5.1. Single and multidimensional approaches

Instead of using a single dimension approach to evaluate research quality, the multiple dimension approach can be applied. For example, the three quality criteria of the Research Assessment Exercise in the United Kingdom are rigor, originality and significance (Bridges, 2009). Ali, Young, and Ali (1996) proposed a 9-item checklist for accessing the quality of a research article, such as originality of research, research methodology, and writing style and interpretation. However, the multiple dimension approach may introduce complexities in representing, processing, and comparing evaluation codes. Therefore, it is better to synthesize the evaluation scores of multiple dimensions into a single score, e.g., using an average or weighted average. After that, the synthesized score can be represented in a reference format and calculated in a similar manner to that explained in the section above.

### 5.2. Comparison with citation count-based indicators

Compared with previous studies on indicators based on citation counts, such as the journal impact factor, the h-index and the like, the proposed method takes the evaluative function of a citation into consideration. Many non-evaluative citations (with the code of "A" or without a code) would be ignored because they only act as links and have no evaluative function. Citation counts are just the number of votes, not the score. A traditional citation has no identifiable, evaluative content, which may be supportive, oppositional, neutral, or an abstention.

Some studies have argued that the number of citations obtained and the eminence of scientists, journals, and institutions are generally closely aligned (e.g., Oppenheim, 1996). However, such studies do not explicitly distinguish the levels of analysis on which the correlations are based (Warner, 2000). A standard way to study the validity of citation count-based indicators is to examine their correlations with peer judgments (Aksnes & Taxt, 2004), because peer review is assumed to be a direct measure of research quality (Moed, 2005). At the macro-level of analysis, comparing research institutions or nations, it is generally agreed that there is a positive correlation between the outcome of peer assessments and citation count-based indicators, which may be because it is harder to conduct peer assessment at a high level of aggregation. However, there are major technical (e.g., erroneous or inaccurate data of citation or affiliation) and methodological problems (e.g., coverage biases in disciplines, countries, and languages) inherent in the current bibliometric system, which make bibliometric analysis inaccurate and unreliable (van Raan, 2005a).

At the meso-level of analysis, comparing journals, research groups, or departments, results showing varying relations, from significantly positive correlation (e.g., Norris & Oppenheim, 2003; Oppenheim, 1997; Saha, Saint, & Christakis, 2003), to weak correlation (Aksnes & Taxt, 2004), to little or no correlation (Goldstein & Maier, 2010; Haddow & Genoni, 2010; Maier, 2006), have been reported. For example, Saha et al. (2003) reported that there was strong correlation between physicians' ratings of the quality of nine general medical journals and the journals' impact factors. Maier (2006) compared impact factors and peer judgment for regional science journals in a European context and found no significant positive correlation between the impact factors and peer judgment. It should be noted that a criticism of some of the studies finding positive correlations is that the criteria used in peer judgments are influenced by citation count-related elements; that is, providing information of citations is an input into the peer evaluation process (Norris & Oppenheim, 2003; Rinia, van Leeuwen, van Vuren, & van Raan, 1998). The original data are not independent, but interrelated. Consequently, such positive correlation results are not convincing.

At the micro-level of analysis, evaluating individual articles or researchers, a number of studies have reported that indicators based on citation counts perform poorly in assessing the quality of individual articles or researchers. For example, West and Mcllwaine (2002) examined correlation between the ratings of two independent expert raters of 79 unsolicited research reports published in the journal *Addiction* in 1997 and the number of citations of these articles. There was no correlation between expert evaluations and the number of citations received by articles. Nieminen et al. (2006) examined whether the statistical and reporting quality of 448 research papers published in 1996 in four psychiatric journals was associated with the number of received citations. They found that the quality of statistical analysis and reporting did not affect the number of citations.

Although some studies report significant positive correlations between citation count-based indicators and peer judgments at the micro-level of analysis, the explained variances are typically small. For instance, Jarwal, Brion, and King (2009) examined the performance of three indicators in predicting the quality of individual research articles as assessed by international reviewers during 2006 and 2007: the journal impact factor and citations per paper, as well as the journal ranking of the Excellence in Research for Australia initiative. Although there were positive correlations between the three indicators and the peer review score, the amount of variance in the peer review score explained by the three indicators was generally less than 20%. They concluded that the three measures are fairly blunt indicators to assess research quality at the individual article level. Bornmann and Leydesdorff (2013) compared rank order correlations between seven citation count-based indicators and the ratings derived from F1000,[1] a post-publication peer review system, using a dataset of 125 papers published in 2008 in the area of cell biology or immunology. The highest indicator explained 20% of the variance in peer rating. They argued that it was reasonable to expect citation count-based indicators to explain only one third of the variance in peer judgments, because these indicators measure one of three aspects of research quality (i.e., impact), while peers are expected to assess all three aspects of research quality (i.e., importance, accuracy, and impact). On the basis of their argument, it might be supposed that more effective indicators/approaches should be developed to measure research quality adequately because peer review has prominent weaknesses.

Aksnes (2006) studied the correlation between citation count-based indicators and authors' own perceptions of scientific contributions of articles and concluded that "highly cited papers could be used as a valid measure of academic scientific excellence, but only at aggregated publication levels. At the individual level, highly cited papers did not necessarily equate to a breakthrough in science or leading edge research" (p. 178), and "at the level of the individual article citations are not… a reliable indicator of a paper's scientific contribution" (p. 182). This conclusion corresponds well with those of Tijssen et al. (2002) and Allen, Jones, Dolby, Lynn, and Walport (2009). Even the founder of the journal impact factor admitted that, "I am like many other authors who feel that their most-cited work is not necessarily their best" (Garfield, 2006, p. 1127).

After reviewing and discussing the findings of the studies on citing behavior for the past four decades, Bornmann and Daniel (2008, p. 69) quoted van Raan's (2005a, p. 134–135) remark as their response to the question of what citation counts measure:

> So undoubtedly the process of citation is a complex one, and it certainly not provides an "ideal" monitor on scientific performance. This is particularly the case at a statistically low aggregation level, e.g. the individual researcher. There is, however, sufficient evidence that these reference motives are not so different or "randomly given" to such an extent that the phenomenon of citation would lose its role as a reliable measure of impact. Therefore, application of citation

[1] http://f1000.com.

> analysis to the entire work, the "oeuvre" of *a group of researchers as a whole over a longer period of time*, does yield in many situations a strong indicator of scientific performance.

This statement has important practical implications. In reality, in most cases when making decisions on awards, appointments, promotions, and funding allocations, decision makers and evaluators assess individual scientists or research works, not groups of scientists or research works as a whole over a long period of time, and they need precise, not approximate, data and results. Indicators based on citation counts have difficulty identifying those highly critically cited works, while $W_{AEI}$ can reveal them. Those indicators do not measure research quality (Pendlebury, 2009), while $W_{AEI}$ does.

Previous indicators based on citation counts give equal positive credit to each citation, regardless of how it is used by the citing author and even if it is cited negatively or neutrally. By using such methods, literature that is ignored effectively loses credits. In contrast with this, the proposed method treats neglected literature in a fair and just way. It is true that an author may not cite all the relevant literature for many reasons, such as visibility, accessibility, and the prejudices of authors and editors (Camacho-Miñano & Núñez-Nickel, 2009; Smith, 1981). Although authors only give evaluations to the cited works they are familiar with, only those references receiving positive evaluations will gain positive credits. Literature that is ignored will not gain credits, but it will not lose any credits either. The status of their $W_{AEI}$ is "unavailable". Non-evaluative citations will not affect the overall scores of $W_{AEI}$.

### 5.3. Comparison with citation motivation analysis

One of the main differences between the proposed method and previous studies on citation motivation analysis is how authors' motivations are collected. Studies of citation motivations are usually based on interviews (Brooks, 1986; Harwood, 2009), questionnaires (Cano, 1989), or manual or automatic citation content analysis (Moravcsik & Murugesan, 1975; Teufel, Siddharthan, & Tidhar, 2006b). The data on citation motivations collected by such methods are not as accurate as would be obtained if the authors designated them at the same time as they cite references, as proposed in the present research, because there are time lags, problems of information recall, misinterpretation by third parties (Harwood, 2009), and so on. "Any method that relies on judgments by persons other than the authors themselves may suffer from reliability problems. This could be a particular problem in citation context and content analyses, which deal with difficult, often complex, and highly specialized subject literatures" (Bornmann & Daniel, 2008, p. 68). It is very difficult, if not impossible, to use mechanized semantic analysis to recognize the quality criteria of rigor, originality, and significance of research publications (Bridges, 2009), as well as identifying the citing authors' motivations. The four-category RES proposed here simplifies the classification of citation functions or motivations constructed by previous studies (Brooks, 1986; Harwood, 2009; Moravcsik & Murugesan, 1975; Teufel et al., 2006a). In Teufel et al.'s (2006a) citation function classification schemes, the differences between different sub-categories within the same categories are subtle. Actually, such differences can be considered as different strengths of attitude on the part of citing authors. The proposed method overcomes the problems of information recall of interview-based or questionnaire-based surveys, unreliability of manual identification, and technical difficulties of automatic recognition of citation motivations and quality evaluations, by requiring citing authors to annotate references clearly. As to the technical issue of the method, it is surely easier for machines to recognize evaluation codes than to understand phrases or words. The present research proposes a quantitative method because it uses a quantitative RES that can indicate the intensity of authors' attitudes, instead of using sub-categories. By this means, the evaluations of citing authors can be aggregated and calculated. It should not be a problem to collect and process the data. In addition, the reference format outlined in the present research is flexible because it allows citing authors to give more than one judgment of each cited work, which reflects the complicated relationships between citing and cited works (Brooks, 1986; Harwood, 2009).

### 5.4. Comparison with publication reputation-based indicators

Given the requirement of academic institutions for faculty to publish articles in high reputation journals, the proposed method might relieve some of the pressure on many scholars, as compared with the method using indicators based on the reputations of publications. Some high quality articles may not be published in high reputation journals for many reasons, e.g., they are controversial, blocked by reviewers, or experience editor bias (Campanario, 2009; Willcocks et al., 2008). A paper published in a high ranking journal may not be better than one in a low reputation journal (Sammarco, 2008). Paul (2008) argued that the quality of papers published in the top ranking journals is not consistent with this expectation, and journal rankings are not a good indicator of research quality. Kacmar and Whitfield (2000) said that it was entirely possible that an article published in a second-tier journal could have more central citations than an article in a top-tier journal. The values of $W_{AEI}$ depend little on where a paper is published. Authors can concentrate on the quality of their work, and spend less time on the selection and submission to journals and conferences.

### 5.5. Comparison with traditional peer review

Unlike traditional peer review that is usually undertaken by only a small group of experts, who have difficulty evaluating fully and fairly a bewildering array of research (Pendlebury, 2009; Taylor, 2011), the method proposed here aggregates all evaluations given by the authors of citing papers. It overcomes the weaknesses of small-scale peer review, and can be regarded as a process of broad peer review. To those who are not familiar with the research works or areas being assessed, $W_{AEI}$ can give assessors more professional opinions. Traditional peer review is criticized for having low inter-reviewer reliability, and lacking fairness and predictive validity (Bornmann, 2011). The likely reasons for this are that the traditional peer review process is closed (a reviewer cannot view another reviewer's opinion), double blind (neither the attributes of reviewers nor reviewees are usually disclosed), and cross-sectional (different peer reviewers examine the same materials at nearly the same period of time). Such a process makes it difficult for researchers to investigate the peer review bias effects because it is hard for them to derive the attributes of reviewers and reviewees. In contrast with this, the process of the proposed method in the present research is open, longitudinal, and even interactive. Different evaluators can examine the same piece of research in different periods and at different points of time. Readers can view previous judgments. In this case, different evaluators' opinions may converge to one or a small number of judgments. In addition, it is easier for researchers to derive the attributes of reviewers and reviewees to test the reviewer bias effect. Therefore, it is possible to achieve higher levels of reliability and help to improve fairness in this open evaluation process.

Previous studies have used citation count-based indicators to verify the predictive validity of peer review (e.g., Bornmann, 2011). However, such an approach is inappropriate. Citation count-based indicators are impact-oriented, while peer review is quality-oriented (Moed, 2005). Research impact only reflects a small portion of research quality (Bornmann & Leydesdorff, 2013). Research quality is the focus of interest in research performance evaluation, while research impact is not. Therefore, it is inappropriate to use citation count-based indicators to test the predictive validity of traditional peer review process. $W_{AEI}$ is a research quality evaluation oriented indicator. It can be considered as a broader peer review process and used to test the predictive validities of traditional peer review process and other partial indicators.

### 5.6. Overall advantages

The proposed method combines the advantages of citation count analysis, citation motivation analysis, and peer review into a single indicator. It is truly transparent. Citations with negative or neutral evaluation or no evaluation at all, are not disregarded, even if they contribute to high citation counts. So the method helps to “differentiate essential from non-essential citations in the production of a scientific paper” (Cano, 1989, p. 284), and therefore differentiates essential from non-essential research works and their contributors in the development of science. The evaluations of references can provide useful information for readers to consider whether they should read the cited works, especially when the readers are looking for cited works that have received complimentary or critical evaluations. This is different from simple popularity votes. Rather, it reflects a progressive process of knowledge discovery and recognition through citing authors' transparent evaluations. Garfield's (1955) main goal when he proposed a bibliographic system for science literature, as well as the concept of impact factor, was to “eliminate the uncritical citation of fraudulent, incomplete, or obsolete data by making it possible for the conscientious scholar to be aware of criticisms of earlier papers” (p. 108). However, it seems citation count-based approaches have not achieved this goal, and may have made the problem worse (Smith, 2006). By suggesting that citing authors give evaluations to their references, the proposed method may help to achieve Garfield's original goal.

Previous indicators have looked at extrinsic correlates of quality, whereas the proposed method assesses research quality. The values of $W_{AEI}$ are sensitive to the evaluations received. They warn people that low quality is bad for the evaluation of research works, scientists, journals/publishers, research groups, and institutions For example, a high quality paper will increase the $W_{AEI}$ of the journal publishing it, while a low quality paper will decrease the journal's $W_{AEI}$. It is hard to maintain a high $W_{AEI}$ because a negative or neutral evaluative citation may reduce $W_{AEI}$. So $W_{AEI}$ may be helpful in discriminating the best research work, scientist, journal/publisher, research group, or institution.

Unlike previous indicators, $W_{AEI}$ is meaningful and comparable between different disciplines and levels of analyses. A uniform evaluation scheme makes comparisons across different disciplines and units of analysis (e.g., comparing a scientist's $W_{AEI}$ and that of a research group) easier.

## 6. Concerns

It should be pointed out that the method proposed here requires the adoption of new referencing regulations, requiring citing authors to give evaluations to their references if they know the field and the cited works well. It is preferable that evaluations of self-citations should be ignored and those of reciprocal citations should be treated with caution.

### 6.1. Malicious evaluation

There is a concern that evaluating others' works is a highly sensitive and subjective activity, and may lead to questionable or unreasonable results, rendering the proposed method unrealistic. However, authors do already explicitly evaluate their sources (Harwood, 2009) and journal editors welcome readers' feedbacks and assessment of the articles they have read (Nallamothu & Lüscher, 2012). Some authors may give positive or negative judgments for malicious reasons, but such behavior by a small number of people will not have a large effect on the overall evaluations. More importantly, the process is totally transparent so that people can see who gives what kinds of evaluations to which works; this should go far in discouraging malicious judgments. The results (of any quantitative analysis for research evaluation) “should be presented openly so that others can understand and check them. Such transparency will help demystify this type of research evaluation” (Pendlebury, 2009, p. 9). It is already the case that researchers being reviewed complain about the anonymous process of traditional peer review, where reviewers are not held to account for their opinions. There have been appeals for an open peer reviewing process, including the selection of the panel (Bence & Oppenheim, 2004). If it were felt to be warranted, a process of mediation could also be introduced using the judgment of additional experts to examine the evaluation. In any case, malicious behaviors have existed for a long time and affect any method.

In reality, many public evaluation systems have already been used to evaluate the performance of products, software, teachers, and universities (e.g., Amazon's customer reviews[2], CNET's software reviews[3], RateYourUni[4], and RateMyProfessors[5]). Online customer product review (also called electronic word-of-mouth) does influence potential buyers' purchase decisions (Chevalier & Mayzlin, 2006; Lee & Lee, 2009). Since 2001, an interactive open access publishing journal using public peer review, *Atmospheric Chemistry and Physics*, has been launched and found to enhance scientific and economic viability and sustainability (Pöschl, 2010). In 2002, a post-publication open peer review online system, named F1000, was launched. It provides experts' reviews and recommendations of top articles in biology and medicine and clinical trials. Nowadays, more and more academic journals let readers give comments to their papers online, e.g., *Nature*. There is a discernible trend in research evaluation to make traditional close peer review processes more public, transparent, and interactive. The proposed method conforms to this trend.

One cannot escape from evaluating or being evaluated. People have given comments to previous works in various forms, e.g., in context, commentary, or review articles. People's knowledge develops over time. Flaws will be found after a period of time, while truths will gain recognition. These developments can be observed from the curves of changing values of $W_{AEI}$. Inappropriate evaluations do happen, even in peer review. Using a larger sample of peer reviewers can help to reduce the influence of bias of some individual peer reviewers (Martin & Irvine, 1983). Indeed, the more evaluations a work receives, the more comprehensive and reasonable opinions people can see.

### 6.2. Effect on the process

Another concern is that the adoption of a new referencing regulation may bring about a radical change in the behavior of authors and journals. References (or footnotes) emerged several hundred years ago (Nicolaisen, 2007). Traditionally, citations were not to evaluate research performance. Since the introduction of the impact factor, indicators based on citation counts have been widely used to evaluate the impact or even quality of research works, scientists, journals, research groups, departments, and institutions. Referencing behavior is complicated (Camacho-Miñano & Núñez-Nickel, 2009), while traditional reference formats are ambiguous. They cannot reflect the delicate relationships between citing and cited works. The ambiguity of traditional reference formats is a reason why using indicators based on raw citation counts to evaluate research performance lack measurement validity. If a citation count is used to evaluate research quality, the traditional reference formats should be improved.

Human behavior follows a route from non-regulation to regulation. When evaluating becomes common or a custom in academic society, regulations requiring authors to give evaluation to their references will be possible. Young generations of researchers and editors, and editors and publishers of electronic publications, are more inclined to accept new methods. Furthermore, there are variations or alternative approaches that can be developed. If it is thought to be too sensitive to have an indication of the evaluations of others' works included in the

[2] http://www.amazon.com/The-Help-Kathryn-Stockett/product-reviews/0425245136/.

[3] http://download.cnet.com/EmEditor-Professional-32-bit/3000-2352_4-10038399.html.

[4] http://www.rateyouruni.com.au.

[5] http://www.ratemyprofessors.com.

citing authors' publications, the citing authors and other appropriate evaluators (e.g., practitioners who have used the finding of those works) can submit the evaluation information to a third party processor dedicated to such evaluations, which may be somewhat similar to the current post-publication open peer review systems, e.g., F1000 and PubZone[6]. In this case, it would not be necessary to change the traditional reference formats. This approach will be explored in further studies.

### 6.3. Limitations

Traditionally, authors have cited others' works in a comparatively casual manner. A potential challenge of the proposed method is that it requires citing authors to give evaluations to their references if they know the field and the cited works well. Obviously, this requirement would take citing authors a considerable amount of extra time and effort to make the assessments. The evaluators are limited to citing authors who are familiar with and willing to give an evaluation of the cited work. Other appropriate evaluators, who know the research well, are not included. In addition, the proposed method is more appropriate for the evaluation of research where the main forms of research outcomes are publications in academic journals or conferences.

## 7. Conclusion

The methodology used to evaluate research performance is one of the cornerstones of academia and it can influence the judgments of decision makers and regulate researchers' conduct of research. It is evolving from an external and somewhat crude approach (e.g., research input-based and publication count-based indicators) to a more internal and finer method (e.g., citation count-based indicators). The current mainstream of research evaluation indicators is based on citation counts, which only show that a piece of research has been used by an academic but do not indicate a citing author's quality-related evaluation of the cited work.

The proposed methods shifts attention toward quality-oriented indicators and make the process of research performance evaluation more interactive, transparent, and public. It should stimulate a number of potential research avenues for further studies. The dimensions included in the construct of research quality used in assessment could be examined and the multi-dimensional approach might be considered. Comparisons could be made of the results obtained from the proposed new method and traditional approaches. Methods for including a broader range of evaluators could be considered. More importantly, thought needs to be given to the attitudes of stakeholders, including editors, publishers, promotion and tenure committees, funding agencies, researchers, practitioners, tax payers, and even general readers toward such an open, interactive research evaluation process.

Although there have been many explorations in measuring research impact, more effort is still needed to achieve the ultimate goal of evaluating research quality. At the current stage, it is best to treat the proposed method as speculation and inspiration rather than as a practical proposal for implementation. It may serve as a preliminary theoretical framework for a quality-oriented approach to research performance evaluation. It is open for discussion and criticism, and can be adjusted, revised, improved, and made more practicable.

## Acknowledgments

The author thanks Professor David Turner for his help in editing.

[6] http://www.pubzone.org/index.do.

## References

Abramo, G., & D'Angelo, C. A. (2011). Evaluating research: From informed peer review to bibliometrics. *Scientometrics*, *87*, 499–514.

Aksnes, D. W. (2006). Citation rates and perceptions of scientific contribution. *Journal of the American Society for Information Science and Technology*, *57*, 169–185.

Aksnes, D. W., & Rip, A. (2009). Researchers' perceptions of citations. *Research Policy*, *38*, 895–905.

Aksnes, D. W., & Taxt, R. E. (2004). Peer reviews and bibliometric indicators: A comparative study at Norwegian University. *Research Evaluation*, *13*, 33–41.

Alberts, B. (2013). Impact factor distortions. *Science*, *340*, 787.

Ali, S. N., Young, H. C., & Ali, N. M. (1996). Determining the quality of publications and research for tenure or promotion decisions: A preliminary checklist to assist. *Library Review*, *45*, 39–53.

Allen, L., Jones, C., Dolby, K., Lynn, D., & Walport, M. (2009). Looking for landmarks: The role of expert review and bibliometric analysis in evaluating scientific publication outputs. *PLoS ONE*, *4*, e5910. Retrieved June 19, 2013, from http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0005910.

Andres, D. L. R., & Wang, M. (2012). Applying psychometric theory and research to developing a continuously distributed approach to making research funding decisions. *Review of General Psychology*, *16*, 298–304.

Archambault, É., & Larivière, V. (2009). History of the journal impact factor: Contingencies and consequences. *Scientometrics*, *79*, 639–653.

Ball, P. (2005). Index aims for fair ranking of scientists. *Nature*, *436*, 900.

Bence, V., & Oppenheim, C. (2004). The influence of peer review on the Research Assessment Exercise. *Journal of Information Science*, *30*, 347–368.

Bergstrom, C. T., West, J. D., & Wiseman, M. A. (2008). The Eigenfactor metrics. *Journal of Neuroscience*, *28*, 11433–11434.

Bollen, J., & Van de Sompel, H. (2008). Usage impact factor: The effects of sample characteristics on usage-based impact metrics. *Journal of the American Society for Information Science and Technology*, *59*, 136–149.

Borgman, C. L., & Furner, J. (2002). Scholarly communication and bibliometrics. *Annual Review of Information Science and Technology*, *36*, 2–72.

Bornmann, L. (2011). Scientific peer review. *Annual Review of Information Science and Technology*, *45*, 199–245.

Bornmann, L., & Daniel, H. D. (2008). What do citation counts measure? A review of studies on citing behavior. *Journal of Documentation*, *64*, 45–80.

Bornmann, L., & Leydesdorff, L. (2013). The validation of advanced bibliometric indicators through peer assessments: A comparative study using data from InCites and F1000. *Journal of Informetrics*, *7*, 286–291.

Bridges, D. (2009). Research quality assessment in education: Impossible science, possible art? *British Educational Research Journal*, *35*, 497–517.

Brooks, T. A. (1986). Evidence of complex citer motivations. *Journal of the American Society for Information Science*, *37*, 34–36.

Camacho-Miñano, M. M., & Núñez-Nickel, M. (2009). The multilayered nature of reference selection. *Journal of the American Society for Information Science and Technology*, *60*, 754–777.

Campanario, J. M. (2009). Rejecting and resisting Nobel class discoveries: Accounts by Nobel Laureates. *Scientometrics*, *81*, 549–565.

Cano, V. (1989). Citation behavior: Classification, utility, and location. *Journal of the American Society for Information Science*, *40*, 284–290.

Casserly, M. F., & Bird, J. E. (2003). Web citation availability: Analysis and implications for scholarship. *College & Research Libraries*, *64*, 300–317.

Chevalier, J. A., & Mayzlin, D. (2006). The effect of word of mouth on sales: Online book reviews. *Journal of Marketing Research*, *43*, 345–354.

Clark, J. G., Au, Y. A., Walz, D. B., & Warren, J. (2011). Assessing researcher publication productivity in the leading information systems journals: A 2005–2009 update. *Communications of the Association for Information Systems*, *29*, 1–48.

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

Darmoni, S. J., Roussel, F., Benichou, J., Thirion, B., & Pinhas, N. (2002). Reading factor: A new bibliometric criterion for managing digital libraries. *Journal of the Medical Library Association*, *90*, 323–327.

Editorial (2005). Ratings games. *Nature*, *436*, 889–890.

Editorial (2006). The impact factor game: It is time to find a better way to assess the scientific literature. *PLoS Medicine*, *3*, e291. Retrieved June 19, 2013, from http://www.plosmedicine.org/article/info:doi/10.1371/journal.pmed.0030291.

Eppler, M. (2006). *Managing information quality: Increasing the value of information in knowledge-intensive products and processes.* New York, NY: Springer-Verlag.

Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. *Scientometrics*, *90*, 891–904.

Frey, B. S., & Rost, K. (2010). Do rankings reflect research quality? *Journal of Applied Economics*, *13*, 1–38.

Garcia, C. E., & Sanz-Menéndez, L. (2005). Competition for funding as an indicator of research competitiveness. *Scientometrics*, *64*, 271–300.

Garfield, E. (1955). Citation indexes for science: A new dimension in documentation through association of ideas. *Science*, *122*, 108–111.

Garfield, E. (1972). Citation analysis as a tool in journal evaluation. *Science*, *178*, 471–479.

Garfield, E. (2006). Commentary: Fifty years of citation indexing. *International Journal of Epidemiology*, *35*, 1127–1128.

Goldstein, H., & Maier, G. (2010). The use and valuation of journals in planning scholarship: Peer assessment versus impact factors. *Journal of Planning Education and Research*, *30*, 66–75.

Gross, P. L. K., & Gross, E. M. (1927). College libraries and chemical education. *Science*, *66*, 385–389.

Haddow, G., & Genoni, P. (2010). Citation analysis and peer ranking of Australian social science journals. *Scientometrics*, *85*, 471–487.

Harwood, N. (2009). An interview-based study of the functions of citations in academic writing across two disciplines. *Journal of Pragmatics*, *41*, 497–518.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, *102*, 16569–16572.

Hussain, S. (2011). Food for thought on the ABS Academic Journal Quality Guide. *Accounting Education*, *20*, 545–559.

Jarwal, S. D., Brion, A. M., & King, M. L. (2009). Measuring research quality using the journal impact factor, citations and "Ranked Journals": Blunt instruments or inspired metrics? *Journal of Higher Education Policy and Management*, *31*, 289–300.

Kacmar, K. M., & Whitfield, J. M. (2000). An additional rating method for journal articles in the field of management. *Organizational Research Methods*, *3*, 392–406.

King, J. (1987). A review of bibliometric and other science indicators and their role in research evaluation. *Journal of Information Science*, *13*, 261–276.

Kostoff, R. N. (1998). The use and misuse of citation analysis in research evaluation. *Scientometrics*, *43*, 27–43.

Lawrence, S., Pennock, D. M., Flake, G. W., Krovetz, R., Coetzee, F. M., Glover, E., Nielsen, F. A., Kruger, A., & Giles, C. L. (2001). Persistence of Web references in scientific research. *Computer*, *34*, 26–31.

Lee, F. S. (2006). The ranking game, class and scholarship in American mainstream economics. *The Australasian Journal of Economics Education*, *3*, 1–41.

Lee, J., & Lee, J. N. (2009). Understanding the product information inference process in electronic word-of-mouth: An objectivity–subjectivity dichotomy perspective. *Information and Management*, *46*, 302–311.

Lehmann, S., Jackson, A. D., & Lautrup, B. E. (2006). Measures for measures. *Nature*, *444*, 1003–1004.

Leydesdorff, L. (2005). Evaluation of research and evolution of science indicators. *Current Science*, *89*, 1510–1517.

Leydesdorff, L. (2008). Caveats for the use of citation indicators in research and journal evaluations. *Journal of the American Society for Information Science and Technology*, *59*, 278–287.

MacRoberts, M. H., & MacRoberts, B. R. (2010). Problems of citation analysis: A study of uncited and seldom-cited influences. *Journal of the American Society for Information Science and Technology*, *61*, 1–12.

Maier, G. (2006). Impact factors and peer judgment: The case of regional science journals. *Scientometrics*, *69*, 651–667.

Martin, B. R. (1996). The use of multiple indicators in the assessment of basic research. *Scientometrics*, *36*, 343–362.

Martin, B. R., & Irvine, J. (1983). Assessing basic research: Some partial indicators of scientific progress in radio astronomy. *Research Policy*, *12*, 61–90.

Moed, H. F. (2005). *Citation analysis in research evaluation*. Dordrecht, Netherlands: Springer Press.

Moed, H. F. (2007). The future of research evaluation rests with an intelligent combination of advanced metrics and transparent peer review. *Science and Public Policy*, *34*, 575–583.

Moravcsik, M. J., & Murugesan, P. (1975). Some results on the function and quality of citation. *Social Studies of Science*, *5*, 86–92.

Nallamothu, B. K., & Lüscher, T. F. (2012). Moving from impact to influence: Measurement and the changing role of medical journals. *European Heart Journal*, *33*, 2892–2896.

Nicolaisen, J. (2007). Citation analysis. *Annual Review of Information Science and Technology*, *41*, 609–641.

Nieminen, P., Carpenter, J., Rucker, G., & Schumacher, M. (2006). The relationship between quality of research and citation frequency. *BMC Medical Research Methodology*, *6*. Retrieved June 19, 2013, from http://www.biomedcentral.com/1471-2288/6/42.

Nightingale, P., & Scott, A. (2007). Peer review and the relevance gap: Ten suggestions for policy-makers. *Science and Public Policy*, *34*, 543–553.

Norris, M., & Oppenheim, C. (2003). Citation counts and the Research Assessment Exercise V: Archaeology and the 2001 RAE. *Journal of Documentation*, *59*, 709–730.

Oppenheim, C. (1996). Do citations count? Citation indexing and the research assessment exercise (RAE). *Serials*, *9*, 155–161.

Oppenheim, C. (1997). The correlation between citation counts and the 1992 research assessment exercise ratings for British research in genetics, anatomy and archaeology. *Journal of Documentation*, *53*, 477–487.

Paul, R. J. (2008). Measuring research quality: The United Kingdom Government's Research Assessment Exercise. *European Journal of Information Systems*, *17*, 324–329.

Pendlebury, D. A. (2009). The use and misuse of journal metrics and other citation indicators. *Archivum Immunologiae et Therapiae Experimentalis*, *57*, 1–11.

Pouris, A. (1989). Evaluating academic science institutions in South Africa. *Journal of the American Society for Information Science*, *40*, 269–272.

Pöschl, U. (2010). Interactive open access publishing and public peer review: The effectiveness of transparency and self-regulation in scientific quality assurance. *IFLA Journal*, *36*, 40–46.

Public Policy Group, L. S. E. (2011). *Maximizing the impacts of your research: A handbook for social scientists (Consultation draft 3)*. London, UK: London School of Economics and Political Science. Retrieved January 15, 2015, from http://www.lse.ac.uk/government/research/resgroups/LSEPublicPolicy/Docs/LSE_Impact_Handbook_April_2011.pdf.

Rinia, E. J., van Leeuwen, Th. N., van Vuren, H. G., & van Raan, A. F. J. (1998). Comparative analysis of a set of bibliometric indicators and central peer review criteria: Evaluation of condensed matter physics in the Netherlands. *Research Policy*, *27*, 95–107.

Saha, S., Saint, S., & Christakis, D. A. (2003). Impact factor: A valid measure of journal quality? *Journal of the Medical Library Association*, *91*, 42–46.

Sammarco, P. W. (2008). Journal visibility, self-citation, and reference limits: Influences on Impact Factor and author performance review. *Ethics in Science and Environmental Politics*, *8*, 121–125.

Seglen, P. O. (1997). Why the impact factor of journals should not be used for evaluating research. *British Medical Journal*, *314*, 498–502.

Smith, L. C. (1981). Citation analysis. *Library Trends*, *30*, 83–106.

Smith, R. (2006). Commentary: The power of the unrelenting impact factor: Is it a force for good or harm? *International Journal of Epidemiology*, *35*, 1129–1130.

Taylor, J. (2011). The assessment of research quality in UK universities: Peer review or metrics? *British Journal of Management*, *22*, 202–217.

Teufel, S., Siddharthan, A., & Tidhar, D. (2006a). An annotation scheme for citation function. In J. Alexandersson, & A. Knott (Eds.), *Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue, Sydney, Australia, July 15–16, 2006* (pp. 80–87). Stroudsburg, PA: ACL.

Teufel, S., Siddharthan, A., & Tidhar, D. (2006b). Automatic classification of citation function. In D. Jurafsky, & E. Gaussier (Eds.), *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP), Sydney, Australia, July 22–23, 2006* (pp. 103–110). Stroudsburg, PA: ACL.

Tijssen, R. J. W., Visser, M. S., & van Leeuwen, T. N. (2002). Benchmarking international scientific excellence: Are highly cited research papers an appropriate frame of reference? *Scientometrics*, *54*, 381–397.

van Raan, A. F. J. (2005a). Fatal attraction: Conceptual and methodological problems in the ranking of universities by bibliometric methods. *Scientometrics*, *62*, 133–143.

van Raan, A. F. J. (2005b). Measurement of central aspects of scientific research: Performance, interdisciplinarity, structure. *Measurement Interdisciplinary Research and Perspectives*, *3*, 1–19.

Warner, J. (2000). A critical review of the application of citation studies to the Research Assessment Exercises. *Journal of Information Science*, *26*, 453–459.

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*, 501–504.

Willcocks, L., Whitley, E. A., & Avgerou, C. (2008). The ranking of top IS journals: A perspective from the London School of Economics. *European Journal of Information Systems*, *17*, 163–168.

Zitt, M., & Bassecoulard, E. (2008). Challenges for scientometric indicators: Data demining, knowledge-flow measurements and diversity issues. *Ethics in Science and Environmental Politics*, *8*, 49–60.

**Zhiqiang Wu** holds a master's in information science from Sun Yat-sen University, China. His current research interests are in the areas of information science, information systems, and management science. He has published in *Scientometrics* and several Chinese journals.