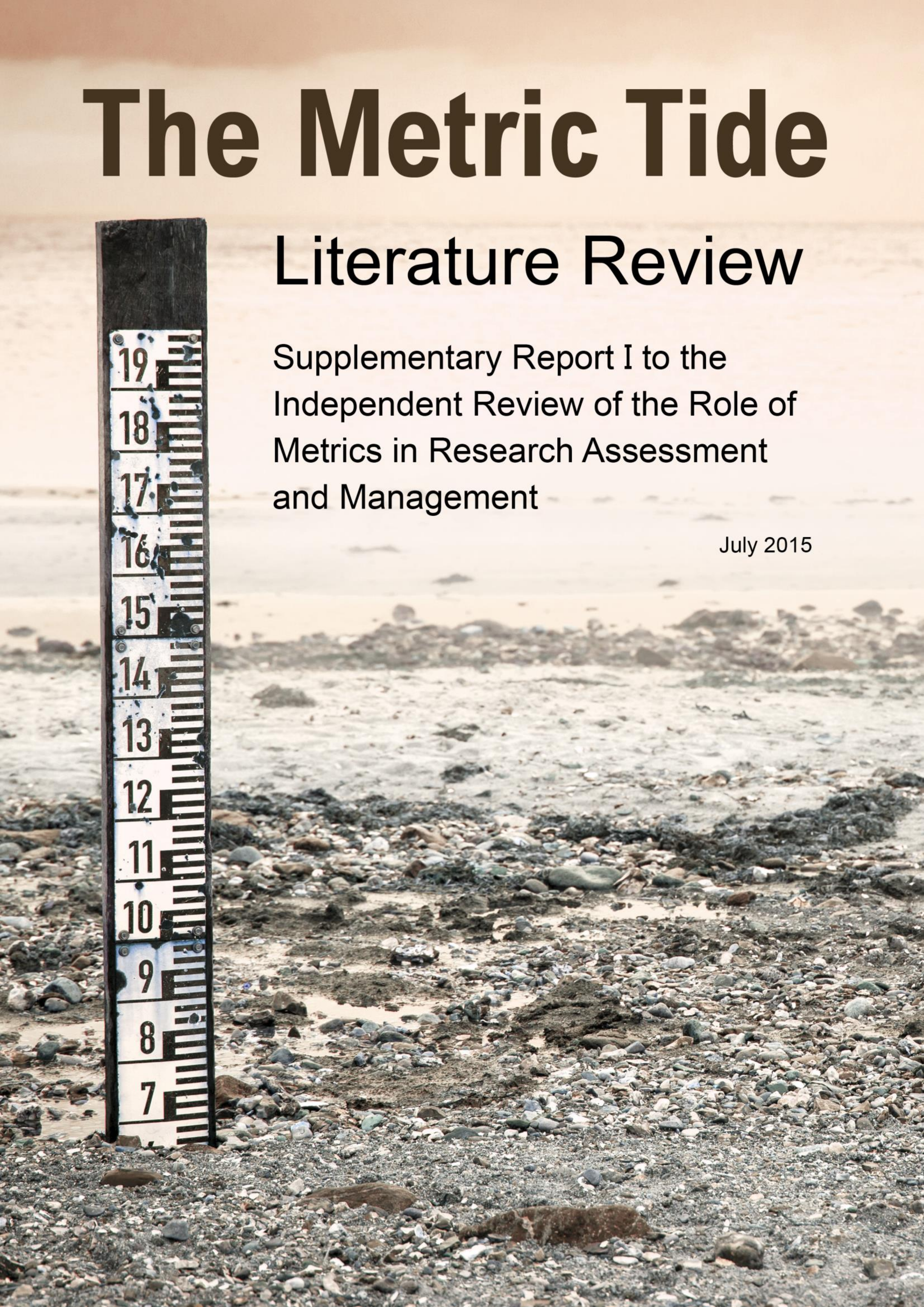
The Metric Tide

Literature Review

Supplementary Report I to the Independent Review of the Role of Metrics in Research Assessment and Management

July 2015



The Metric Tide

Literature Review

Supplementary Report I to the Independent Review of the Role of Metrics in Research Assessment and Management

July 2015

Authors
Paul Wouters*, Mike Thelwall**, Kayvan Kousha**, Ludo Waltman*, Sarah de Rijcke*, Alex Rushforth*, and Thomas Franssen*

*Centre for Science and Technology Studies (CWTS), Leiden University, Netherlands.
**Statistical Cybermetrics Research Group, University of Wolverhampton, UK.

Cite as
Wouters, P. et al. (2015). *The Metric Tide: Literature Review (Supplementary Report I to the Independent Review of the Role of Metrics in Research Assessment and Management)*. HEFCE. DOI: 10.13140/RG.2.1.5066.3520

ISBN: 1902369280
DOI: 10.13140/RG.2.1.5066.3520
Twitter: #HEFCEmetrics

Contents

Executive Summary

This document provides a literature review of academic research about a range of indicators that may be useful in research evaluations, including the next Research Excellence Framework (REF). It was commissioned by the Higher Education Funding Council for England (HEFCE) as part of the independent review of the role of metrics in research assessment that commenced in 2014[1]. The purpose of the document is to help to ensure that the recommendations of the independent review are informed by relevant academic research. The literature review is focussed on three overarching areas which are: bibliometrics and the use of indicators; peer review and bibliometrics; and alternative metrics.

Bibliometrics and the use of indicators

The literature review explores several of the key issues and considerations relating to bibliometrics analysis, as listed below

Bibliographic databases

The three most important multidisciplinary bibliographic databases are Web of Science (WoS), Scopus, and Google Scholar (GS). Coverage of the scientific literature across subject disciplines varies significantly between these and other databases. GS is generally found to outperform both WoS and Scopus in terms of its coverage of the scientific literature. However, it is often criticised for its lack of quality control and transparency and currently has the disadvantage of being difficult to use for large-scale bibliometric analyses. The Social Sciences and Humanities (SSH) create special challenges for bibliometric analyses, primarily as books and national journals can play an important role in these subject areas and such publications are less often indexed in bibliographic databases. Bibliometric analyses in computer science and engineering involve similar difficulties as many publications appear in conference proceedings which tend to be less well covered by bibliographic databases than is the journal literature. Another problem related to proceedings literature is that the same work may be published multiple times, for instance first in conference proceedings and then in a journal.

Basic citation impact indicators

A large number of citation impact indicators have been proposed in the literature. Most of these can be seen as variants or extensions of a limited set of basic indicators. These basic citation impact indicators are the total and the average number of citations of the publications of a research unit (e.g.

[1] http://www.hefce.ac.uk/rsrch/metrics/

of an individual researcher, a research group, or a research institution), the number and the proportion of highly cited publications of a research unit, and a research unit's h-index. The benefits and disadvantages of these approaches are reviewed.

Exclusion of specific types of publications and citations

Bibliometricians are required to decide which types of publications and citations should be excluded within their analysis. Whilst documents classified as research articles are usually included, some, such as 'editorial material', 'letter', and 'review', are more often excluded, as are some non-English language publications or publications in national journals, which can create biases. Self-citations are sometimes also excluded, though there is no consensus on the optimum approach in the literature.

Normalisation of citation impact indicators

There is agreement in the literature that citation counts of publications from different fields should not be directly compared with each other. This is because there are large differences among fields in the average number of citations per publication. Researchers have proposed various approaches to normalise citation impact indicators for field differences, some being based on average citations and others on highly cited publications. However, a key issue in the calculation of normalised citation impact indicators is the way in which the concept of a scientific field is operationalised, and there are differences in opinion as to how this should be undertaken, for instance, through predefined database fields, disciplinary classification systems, or sophisticated computer algorithms to define fields or with citing-side normalisation approaches which do not define fields explicitly.

Credit allocation in the case of multi-author publications

The average number of authors of publications in the scientific literature keeps increasing, indicating a trend towards greater collaboration in science, which makes it increasingly difficult to properly allocate the credit for a publication to the individual authors. The most common approach is to allocate the full credit for a publication to each individual author, known as full counting. However, this approach has an inflationary effect, which is sometimes considered undesirable. A number of alternative credit allocation approaches have therefore been proposed in the literature, namely: the fractional counting approach; full allocation of publication credit to the first author; full allocation of publication credit to the corresponding author; and allocation of publication credit to the individual authors in a weighted manner, with the first author receiving the largest share of the credit.

Indicators of the citation impact of journals

The best-known indicator of the citation impact of journals is the impact factor. There is considerable debate about the impact factor, both regarding the way in which it is calculated and regarding the way in which it is used in research assessment contexts. Various improvements of and alternatives to the impact factor have been proposed in the literature. These include: (1) Taking into account citations during a longer time period, possibly adjusted to the specific citation characteristics of a journal, or considering the median instead of the average number of citations of the publications in a journal, or calculating an h-index for journals as an alternative or complement to the impact factor; (2) Normalising for differences in citation characteristics among fields, for instance as implemented in the SNIP indicator available in Scopus; (3) Attributing more weight to citations from high-impact sources, for instance as implemented in the eigenfactor indicator included in the WoS-based journal citation reports and the SJR indicator included in Scopus. The impact factor and other citation impact indicators for journals are often used not only in the assessment of journals as a whole but also in the assessment of individual publications in a journal. Journal-level indicators then serve as a substitute for publication-level citation statistics, a practice which is rejected by many bibliometricians, as the distribution of citations over the publications in a journal is highly skewed, which means that the impact factor and other journal-level indicators are not representative of the citation impact of a typical publication in a journal.

Main research strands on indicator effects

The literature review also summarises the main research strands on indicator effects, noting the wide-ranging set of literatures that focus on the governance of science at large and the multi-disciplinary body of work that portrays the rise of performance measurement in academic settings as part of a broader upsurge of accountability measures in public institutions from the 1980s onward. Indicators are positioned as tools that drive competition, instrumentality and privatisation strategies and help steer academic institutions and researchers towards becoming more like market-oriented actors. Science policy studies focus mainly on existing and new types of formal assessment tools, and on providing methods for research evaluation. They concentrate on formalised national evaluation systems, in which citation-based indicators do not play a prominent role. The sociology of science views scientific quality control as a thoroughly social and organisational process. This field analyses accountability measures in the light of transformations to institutional and organisational dynamics of science and innovation. Most studies do not deal concretely with the effects of indicator uses, however. The review discusses a number of strands in some depth, including:

- **Strategic behaviour and goal displacement.** Several studies indicate that funding and evaluation regimes have led to goal displacement in a number of countries (a process in which scoring high on performance measures becomes a goal in itself, rather than a means of measuring whether a certain performance level has been attained). For instance, there are indications that the RAE affected researchers' choices for particular outlets.

- **Effects on interdisciplinarity**. Evidence for the impact of disciplinary assessments such as the REF on interdisciplinary research varies. Studies show that negative effects such as goal displacement do occur, but the dynamics are most likely discipline-specific.

- **Task reduction.** Studies confirm the abandonment of particular types of work (e.g. teaching, outreach), and a focus on particular publication forms (international, peer-reviewed journal articles) and types of research topics (the mainstream). Again, the known effects vary per discipline and the evidence is limited.

- **Effects on institutions.** The rise of formal evaluation may appear isomorphic, but existing research does not reveal uniformity in institutional-level transformations of governance structures. Some studies do find an alignment between institutional measures and the standardised criteria of formal evaluation agencies (e.g. mirroring institutional hiring or promotion strategies with criteria set by funders). Some analysts explain that the legitimacy of indicators in generating and ordering strategic information on performance rests for a large part on the authority they carry within institutional environments – relatively independent of any first-order (in)-accuracies. The functioning of bibliometric tools in 'reducing complexity' is frequently cited as a reason for their widespread appeal among policymakers and research managers.

- **Effects on knowledge production.** Recent studies find a reification of evaluative metrics in research management and decision-making contexts, as formal and informal standards against which research activities are assessed, in some fields. Some analyses point to a discrepancy between the importance of indicators in evaluation practices according to academics and their own judgment of the accuracy of certain measures. However, the sophistication of academics' understanding of the (dis-)advantages of performance measures should not be underestimated. The use of performance indicators and advanced bibliometric information may also influence the conditions under which research agendas are developed (e.g. scientometric information may potentially play a role in democratising the process of agenda-setting). Again, the evidence is fragmented and not complete.

Peer review and bibliometrics

Peer review is a general umbrella term for a host of expert-based review practices that show considerable variation, including journal review of manuscripts, peer review of applications for funding and career promotions, and national peer review-based research assessments.

The results of peer review-based decisions generally show positive correlations with selected bibliometric performance data, but this varies depending on which forms of peer review and which specific dimensions of peer review are being related to which bibliometric indicators. It is also important to define exactly how these bibliometric indicators are being measured and on the basis of which data sets. Bibliometric measures ought not by definition to be seen as the objective benchmark against which peer review is to be measured.

The literature on the relationship between peer review and bibliometrics has yet to develop a common methodology. As a result, different outcomes of studies with respect to the strength of the correlation between peer review and bibliometric measurement may be caused by different research designs. Many studies of the relationship between bibliometric performance and funding decisions that report a positive correlation are plagued by circular reasoning: the better citation performance of funded researchers may very well be the result of this funding.

The popular view that citation rate is a measure of scientific quality is not supported by the bibliometric expert community. Bibliometricians generally see the citation rate as a proxy measure of scientific impact or of impact on the relevant scientific communities. This is one of the dimensions of scientific or scholarly quality. Quality is seen as a multidimensional concept that cannot be captured by any one indicator. Moreover, which dimension of quality should be prioritised in research assessments may vary by field and by research mission.

The literature shows varying strengths of correlation between bibliometric indicators and peer review assessment. Correlation strengths vary between fields within the natural sciences, the social sciences, and the humanities, and it may even vary within fields. In general, the correlation between bibliometrics and peer review is weaker in most fields in the humanities, the applied fields, the technical sciences, and the social sciences. This is partly caused by less coverage in the citation databases, but also by varying citation and publication cultures.

Peer review and bibliometric data are not completely independent. Citation data are in the end based on scientists who cite or do not cite particular publications. The same communities are the source of the peer review data. Although the meaning of the citation cannot be deduced from the role of the literature reference, it does explain the strongly to moderately positive correlation between peer review and bibliometrics. In addition, peer review decisions may have been influenced by prior knowledge of bibliometric data. This interaction may have increased due to the large-scale availability of bibliometric data and indicators.

The strength of peer review is also its weakness. In the context of national research assessments, the literature identifies the inevitable selectivity of post-publication peer review as a possible problem of exclusively peer review-based evaluation. This may be an area where publication and bibliometric data may add value.

The literature does not currently support the idea of replacing peer review by bibliometrics in the REF. First of all, the existence of strong correlations between bibliometrics and peer review is a necessary but not sufficient condition for replacing peer review by bibliometrics. It depends on which parts of the role of peer review in the REF one wishes to prioritise. Only if the exclusive goal of the REF were to be the distribution of research funding among the universities, could one consider replacing the REF by advanced bibliometric analysis for a number of fields. (However, one should then also consider renaming the funding stream.) Second, the studies that confirmed strong correlations still showed strong variation of ranking results at the level of research institutions. At higher levels of aggregation, the correlation generally becomes stronger. Third, not all fields show strong correlations between bibliometrics and peer review data.

The literature does support the idea of supplementing peer review by bibliometrics (informed peer review). Currently, this concept has not yet been formally operationalised. Bibliometric data may counter specific weaknesses of peer review (its selectivity, particular forms of bias, etc.). Experiments with forms of informed peer review therefore seem the best way forward. Metrics may also help to open up the process of disciplinary peer review to include criteria for societal impact of research.

Alternative metrics for research evaluation

More than a decade ago, the rise of new ways for scholars to write, communicate and publish research via electronic media led to calls for novel indicators for electronic scholarly communication. In response, alternative indicators have been developed to capture evidence of types of impact which include *web citations* in digitised scholarly documents (e.g. eprints, books, science blogs or clinical guidelines) or, more recently, *altmetrics* derived from social media sources (e.g. social bookmarks, comments, ratings, microblog posts). Scholars nowadays may also produce and use non-refereed academic outputs, such as multimedia products, datasets and software. It is important to estimate the impact of these non-standard outputs too, if possible, and new usage-based indicators would be needed for this.

In summary, alternative metrics may be helpful when evaluators, funders or even national research assessments need to know ‘all kinds of social, economic and cultural benefits and impacts beyond academia’ (REF, 2011, p. 4) as well as non-standard impacts inside academia. The literature review provides an overview of the findings of research into many different types of alternative metrics. It also discusses the use of associated indicators for assessing the impact of articles, books and other academic outputs (e.g. science videos, datasets and software). Summary guidelines are also given of the potential advantages and limitations of these alternative metrics over traditional bibliometric indicators.

The conclusions for this element of the review indicate that a wide range of indicators derived from the web for scholars or their outputs are related to scholarly activities in some way because they correlate positively and significantly with citation counts. In many cases these metrics can also be harvested on a large scale in an automated way with a high degree of accuracy. Nevertheless, most are easy to spam and nearly all are susceptible to spam to some extent. Moreover, whilst a few seem to reflect types of impact that are different from that of traditional citations, none except Google patent citations and clinical guideline citations clearly reflect wider societal impact. In addition, many of them seem to be too rare to help to distinguish between the impacts of typical publications, but could be useful to give evidence of the impact of the small minority of high impact articles. Therefore, despite the considerable body of mostly positive empirical evidence, with some exceptions alternative metrics do not seem to be useful to capture wider social impact and do not seem to be robust enough to be routinely used for evaluations in which it is in the interest of stakeholders to manipulate the results. In other words, alternative metrics do not seem to be suitable as a management tool with any kind of objective to control researchers. Even if no manipulation took place, which seems unlikely, the results would be suspected to be affected by manipulation and in the worst case the results would be extensively manipulated and scientists would waste their time and money on this manipulation.

1. Bibliometrics and the use of citation indicators in research assessment

The HEFCE Steering Group on the role of metrics in research evaluation in general, and in future instalments of the UK Research Excellence Framework (REF) in particular, will consider how well metrics can be used across different academic disciplines to assess the excellence of research undertaken in the higher education sector. This part of the literature review focuses on the state of the art with respect to bibliometric indicators. We used the Call for Evidence as our frame of reference to define the scope of the literature involved. In this call a number of topics have been raised:

1. identifying useful metrics for research assessment;
2. how should metrics be used in research assessment?;
3. 'gaming' and strategic use of metrics.

In addition, the call requests respondents to include relevant evidence and examples from outside the UK. This literature review is also a contribution to this last demand since it is based on the international literature, irrespective of the country or discipline involved.

This review addresses the three topics in the Call for Evidence in the following way:

The use of bibliometrics in research evaluation ultimately depends on the coverage and quality of the databases that are used. Currently, three international bibliometric databases are dominant (Web of Science (WoS), Scopus, and Google Scholar (GS)), but their properties are not always well understood. In the first part of this review, we therefore present and discuss the current literature on these three databases, their comparative performance in the context of research evaluation, and the implications for performance analysis based on these databases. This part of the review addresses topics 1 and 2 of the Call for Evidence.

In the second part of this review, we present an analysis of the most important types of bibliometric indicators for research assessment. Scientometrics is a field that has exploded in the size of its literature as well as in the diversity of its practitioners. Whereas in the early 1960s only a small group of sociologists and information scientists were interested in scientometrics (grouped around Eugene Garfield, Robert Merton, and Derek de Solla Price respectively), nowadays the field has hundreds of full-time researchers, and many more part-time users of scientometric data and indicators. As a result, the literature on indicators is strongly biased towards the invention and study of new indicators. This literature review is the first review that has created a more fundamental typology of different classes of indicators and their properties. We believe that this will not only be useful in the context of the work of the HEFCE Steering Group, it may also be an important road map for bibliometricians and citizen scientometricians who try to find their way in this massive and widely spread literature. This part of the review also addresses topics 1 and 2 of the Call for Evidence.

In the third part of the review, we review the international literature on evaluation practices, 'gaming' of indicators and evaluation exercises, and strategic responses by scientific communities and others to requirements in research assessments. This literature is of a different character than the literature covered in the first two parts of this review.

First of all, it is published in different media. The technical bibliometrics literature on databases and indicators has been published in international journals that are well covered by WoS and Scopus. In contrast, the literature on evaluation practices is spread over journals, books, edited volumes (which sometimes play the role of journals but are less accessible), and reports and other forms of grey literature.

Secondly, its spread over disciplines differs substantially. The literature about bibliometric data and indicators is concentrated in one sub-field within the field of information and library sciences: bibliometrics. This is a well-organised field with a host of journals, three leading conference series, and a tightly linked core of full-time researchers. The literature on evaluation practices, in contrast, is spread over a relatively large number of social science fields, among which are: sociology of science, innovation studies, library and information science, higher education studies, sociology of evaluation, medical sociology, evaluation studies and designs, economics and business studies, social psychology, standards and indicator research, complexity science, science policy studies, research management and innovation, political science, and governance studies.

Thirdly, the epistemic nature of the two types of literature is different. The bibliometric literature consists of technical reporting about data and data analyses, the construction of indicators, or model and simulation experiments. The literature on evaluation practices is of a more qualitative and heterogeneous character. It includes: surveys among researchers about their views and experiences with evaluations; formal policy analysis of principal-agent relationships; ethnographic studies of evaluation in action (a recently emerging body of work); reflexive studies of evaluation researchers; cultural critiques of the evaluation society or new public management theories; and last but not least reflections by scholars and scientists on their experiences with the latest round of assessments in higher education. As a result, the nature of the evidence presented in the literature varies considerably. Although the size of this literature is smaller than the vast body of work on bibliometric indicators, its more heterogeneous character presents specific intellectual challenges in its incorporation in this review. We present the main strands in the literature, with a focus on empirical materials about possible effects of indicators and strategic responses by the scientific communities.

1.1. Bibliographic databases

The three most important databases available for performing bibliometric analyses are WoS, Scopus, and GS. There are more databases available, but these usually cover only a limited number of scientific fields. We start by summarising some key features of WoS, Scopus, and GS.

WoS is a subscription-based database that comprises a number of citation indices. The best-known citation indices are the Science Citation Index Expanded, the Social Sciences Citation Index, and the Arts & Humanities Citation Index. These citation indices cover journals and book series. Nowadays, WoS also offers a Conference Proceedings Citation Index and a Book Citation Index, covering conference proceedings and books. WoS was originally owned by the Institute for Scientific Information. Its current owner is Thomson Reuters. The use of WoS in bibliometric analyses has a long history, and bibliometricians have therefore studied the characteristics of this database in significant detail. For instance, Moed (2005a, Chapter 7) and Larsen and Von Ins (2010) analyse the coverage of WoS, Michels and Schmoch (2012) investigate the growth of WoS, Harzing (2013a) studies the document type classification of WoS, and García-Pérez (2011) draws attention to the issue of incorrect citation relations in WoS.

Like WoS, Scopus is a subscription-based database. In addition to journals, Scopus also covers trade publications, book series, conference proceedings, and books. Scopus is owned by Elsevier and was launched in 2004.

GS was also launched in 2004. It indexes scholarly literature that is available on the web. This includes not only publications in journals and conference proceedings, but also for instance books, theses, preprints, and technical reports. GS is made freely available by Google. Little is known about the coverage of GS. For instance, there is no list available of sources that are covered by GS. In a recent study by Khabsa and Giles (2014), it is estimated that GS indexes about 100 million documents, representing almost 87% of all English-language scholarly documents available on the web.

Most institutions with a subscription to WoS or Scopus have access to these databases through a web interface. The WoS and Scopus web interfaces can be used for performing simple bibliometric analyses at a relatively small scale. Advanced bibliometric analyses at a larger scale require direct access to the full WoS or Scopus database, without the restrictions imposed by a web interface. Professional bibliometric centres often have direct access to the full WoS or Scopus database. An alternative way of performing advanced bibliometric analyses is the use of specialised web-based tools such as InCites and SciVal. InCites is provided by Thomson Reuters based on WoS, and SciVal is provided by Elsevier based on Scopus. Performing large-scale bibliometric analyses using GS is more difficult, because the only way to access GS is through its web interface. It is not possible to get direct access to the full GS database. Bibliometric analyses based on GS are sometimes performed using a software tool called Publish or Perish (Harzing, 2010).[2]

Below, we provide an overview of the literature on WoS, Scopus, and GS. We emphasise that all three databases are in continuous development (e.g. Chen, 2010; De Winter et al., 2014; Harzing, 2013b, 2014; Michels & Schmoch, 2012). Results reported in the literature, especially in less recent work, may therefore not be fully up-to-date anymore.

1.1.1. Comparing WoS and Scopus

The most comprehensive comparison between WoS and Scopus is reported by Moed and Visser (2008). By matching publications in Scopus with publications in WoS, they establish that 97% of all publications from 2005 covered by WoS are also covered by Scopus. Hence, WoS can almost be considered a perfect subset of Scopus. Looking at publications submitted to the 2001 Research Assessment Exercise (RAE), Scopus coverage turns out to be better than WoS coverage especially in the subject group 'Subjects Allied to Health' and to a lesser degree in the subject group 'Engineering & Computer Science'. Moed and Visser (2008) note that when Scopus is used in a bibliometric analysis it may be preferable to work with a subset of all Scopus data instead of the entire database. As reported by López-Illescas et al. (2008, 2009) based on an analysis in the field of oncology, journals covered by Scopus and not covered by WoS tend to have a low citation impact and tend to be more nationally oriented. Including these journals in a bibliometric analysis may significantly reduce the average citation impact of certain countries.

The observation that Scopus has a better coverage than WoS is made in various other studies as well. In an analysis of researchers in the field of human-computer interaction, Meho and Rogers (2008) observe that Scopus has a better coverage of conference proceedings than WoS. Gavel and Iselid (2008) find that Scopus has better journal coverage than WoS especially in science, technology, and medicine. Norris and Oppenheim (2007) observe that Scopus has a better coverage than WoS of social science publications submitted to the 2001 RAE. In an analysis of Slovenian publications, Bartol et al. (2014) report that Scopus has a better coverage than WoS in the social sciences, humanities, and engineering and technology. In a study of publications of two Portuguese universities, Vieira and Gomes (2009) observe publications that are covered by Scopus and not by WoS, but the reverse situation is found as well. It is also observed in the literature that citation counts tend to be higher in Scopus than in WoS (e.g. Haddow & Genoni, 2010; Kulkarni et al., 2009; Torres-Salinas et al., 2009).

[2] Publish or Perish can be downloaded from www.harzing.com/pop.htm. It is free for personal non-profit use.

Regarding the sensitivity of citation analyses to the choice between WoS and Scopus, results reported in the literature are somewhat mixed. Torres-Salinas et al. (2009) report that WoS and Scopus yield similar results for rankings of university departments. In an analysis in the field of information studies, Meho and Sugimoto (2009) observe that for smaller entities (e.g. journals, conference proceedings, and institutions) results based on WoS and Scopus are considerably different while for larger entities (e.g. research domains and countries) very similar results are obtained. This is in line with Archambault et al. (2009), who show that at the country level results based on WoS and Scopus are highly correlated.

1.1.2. Comparing GS with WoS and Scopus

There are a substantial number of studies in which GS is compared with WoS and sometimes also with Scopus. A number of studies report that GS outperforms WoS and Scopus in terms of coverage of publications. Meho and Yang (2007) analyse publications of library and information science researchers and report that, in comparison with WoS and Scopus, GS stands out in its coverage of conference proceedings and non-English language journals. Mingers and Lipitakis (2010) find that in the field of business and management GS has a much better coverage than WoS, and they therefore conclude that WoS should not be used for measuring the impact of business and management research. Very similar observations are made by Amara and Landry (2012). Walters (2007) reports that GS has a substantially better coverage than WoS in the field of later-life migration. Franceschet (2010a) compares WoS and GS in an analysis of computer science researchers and finds that GS identifies many more publications and citations than WoS. A similar result is obtained by García-Pérez (2010) in the field of psychology. Kousha and Thelwall (2008b) study the sources of citations that are counted by GS but not by WoS. They find that 70% of these citations originate from full-text scholarly sources available on the web.

On the other hand, there are studies indicating that the coverage of GS is not consistently better than the coverage of WoS and Scopus. The analysis of Bar-Ilan (2008b) suggests that GS has a better coverage than WoS and Scopus in computer science and mathematics, but a worse coverage in high energy physics. Bornmann et al. (2009) report coverage problems of GS in the field of chemistry. They conclude that WoS and Scopus are more suitable than GS for research evaluation in chemistry. Bakkalbasi et al. (2006) compare WoS, Scopus, and GS in the fields of oncology and condensed matter physics and indicate that none of the three databases consistently outperforms the others. A similar conclusion is reached by Kulkarni et al. (2009) based on an analysis of publications in general medical journals. Mikki (2010) presents a comparison of WoS and GS in the field of earth sciences and also reports that neither database has a consistently better performance than the other. It should be noted that a number of studies indicate substantial improvements in the coverage of GS over time (Chen, 2010; De Winter et al., 2014; Harzing, 2013b, 2014). This suggests that perhaps the results of earlier studies reporting coverage problems of GS may not be relevant anymore.

A different perspective on the comparison of GS with WoS and Scopus is offered by Li et al. (2010). These authors use WoS, Scopus, and GS to calculate bibliometric indicators for a number of library and information science researchers, and they then correlate the indicators with judgments provided by experts. The three databases turn out to yield broadly similar results. Indicators calculated based on Scopus are most strongly correlated with expert judgment, while WoS-based indicators have the weakest correlation, but the differences are very small.

Various studies also investigate specific problems with GS. A general conclusion from the literature is that there is a lack of quality control in GS. There are many inaccuracies and errors. Jacsó (2006) for instance discusses the problem of incorrect citation counts in GS. The possibility of manipulating citation counts in GS is discussed by Beel and Gipp (2010a), Labbé (2010), and López-Cózar et al. (2014). GS is also criticised for its lack of transparency (e.g. Wouters & Costas, 2012). It is unclear what is covered by GS and what is not. Researchers also point out that cleaning GS data can be very time consuming (Li et al., 2010; Meho & Yang, 2007).

1.1.3. Social sciences and humanities

Social sciences and humanities (SSH) research differs from research in the sciences in a number of fundamental ways. This is discussed in detail in literature reviews provided by Hicks (1999), Nederhof (2006), and Huang and Chang (2008). Nederhof (2006) for instance lists the following key differences between SSH research and research in the sciences:

- SSH research has a stronger national and regional orientation.
- SSH research is published less in journals and more in books.
- SSH research has a slower pace of theoretical development.
- SSH research is less collaborative.
- SSH research is directed more at a non-scholarly public.

As pointed out by Hicks (1999) and Nederhof (2006), because of the relatively strong national and regional orientation of SSH research, the coverage of SSH publications in WoS is limited. Many national and regional SSH journals are not covered by WoS. The significant role played by books in SSH research also contributes to the limited WoS coverage of SSH publications. Until recently, books were not covered at all by WoS.

The difficulties caused by the national and regional orientation of SSH research are emphasised by Archambault et al. (2006). They claim that WoS has a 20-25% overrepresentation of English language SSH journals. On the other hand, the difficulties caused by book publishing may diminish over time.

Larivière et al. (2006) observe that journals play an increasingly important role in the social sciences (but not in the humanities). Also, WoS nowadays includes a Book Citation Index. This may make it possible to include books in bibliometric analyses, although Gorraiz et al. (2013) conclude that the Book Citation Index at the moment should not yet be used for bibliometric purposes.

Further insight into the WoS coverage of SSH literature is provided by studies in which the complete SSH publication output of a country or region is compared with the output that is covered by WoS. Such studies are reported by Larivière and Macaluso (2011) for the province of Québec in Canada, by Engels et al. (2012) for the region of Flanders in Belgium, and by Sivertsen and Larsen (2012) for Norway. Larivière and Macaluso (2011) study the Érudit database, which is a database of journals from Québec, and report that in comparison with WoS this database includes about 30% more SSH publications from French-speaking universities in Québec. Based on an analysis of the VABB-SHW database, which is a database of publications authored by SSH researchers in Flanders, Engels et al. (2012) conclude that SSH researchers in Flanders increasingly publish their work in English, often in WoS covered journals, but they also report that there is no shift away from book publishing. The main observation made by Sivertsen and Larsen (2012), based on data on SSH publications from Norway, is that book publishing and domestic journal publishing show a concentration of many publications in a limited number of publication channels, which suggests that there are promising opportunities for obtaining a more comprehensive coverage of SSH literature. A comparison between the databases used in Flanders and Norway is presented by Ossenblok et al. (2012), who conclude that SSH researchers in Flanders display a stronger tendency to publish in WoS covered journals than Norwegian SSH researchers.

1.1.4. Conference proceedings

In certain fields, publications in conference proceedings play an important role. This is especially the case in computer science and engineering, and to some extent also in the social sciences, as shown by Glänzel et al. (2006b) and Lisée et al. (2008). However, including conference proceedings publications in a bibliometric analysis is difficult for a number of reasons. Below, we discuss two important difficulties.

The first difficulty is that little is known about the coverage of conference proceedings in WoS, Scopus, and GS. There is almost no work in which the three databases are compared. Exceptions are the studies by Meho and Rogers (2008) and Meho and Yang (2007), both of which have already been mentioned above. Meho and Rogers (2008) report that in the field of human-computer interaction Scopus has a better coverage of conference proceedings than WoS. Meho and Yang (2007) find that in the field of library and information science GS outperforms WoS and Scopus in terms of its coverage of conference proceedings. Another study of the coverage of conference proceedings is reported by Michels and Fu (2014), but this study considers only the WoS database. Michels and Fu observe gaps in the coverage of important conferences in WoS.

The second difficulty in the use of conference proceedings publications in a bibliometric analysis relates to the issue of double counting of work that is published both in conference proceedings and in a journal. This issue is analysed by Bar-Ilan (2010) and Michels and Fu (2014). As pointed out by Bar-Ilan, double counting creates various problems. Most importantly, publication counts increase in an artificial way as a consequence of double counting, while citation counts per publication are likely to decrease.

1.2. Citation impact indicators

The literature on citation impact indicators is large. Lots of proposals on new indicators have been made in the literature. Providing a comprehensive overview of all proposals is hardly possible. Our aim therefore is to focus on the main ideas suggested in the literature. Indicators that have received a lot of attention in the literature or that play an important role in practical applications will be discussed. Many other indicators will not be included in the discussion or will be mentioned only briefly.

Other overviews of citation impact indicators are provided by Vinkler (2010) and Wildgaard et al. (2014). The monograph by Vinkler (2010) offers a systematic overview of scientometric indicators for research evaluation. Wildgaard et al. (2014) present a review of the literature on bibliometric indicators for measuring the performance of individual researchers.

1.2.1. Basic citation impact indicators

To organise the discussion on citation impact indicators, we start by distinguishing a number of very basic indicators. These basic indicators are important because almost all indicators proposed in the literature can be seen as variants or extensions of these basic indicators.

We assume that we know how many citations have been received by each publication of a research unit[3]. As discussed in Section 1.1, the number of publications of a research unit and the number of citations of these publications are likely to be different in different databases. In the present section, we simply assume that one particular database is used and we work with the publication and citation counts provided by that database. The number of publications of a research unit also depends on the

[3] We use ‘research unit’ as a general term that may for instance refer to individual researchers, research groups, departments, research institutions, or journals. In most of the discussion, we do not need to distinguish between these different types of research units.

time period within which publications are counted. Likewise, the number of citations of a publication depends on the time period within which citations are counted. Instead of simply counting all publications and citations, one may choose to count only publications from for instance the past 10 years and only citations received within for instance the first five years after the appearance of a publication. In this section, we do not discuss these choices in more detail. We simply assume that we work with a given set of publications and for each publication a given number of citations.

Table 1 lists five basic citation impact indicators:

- *Total number of citations*. The total number of citations of the publications of a research unit. As an example, consider a research unit with five publications, which have 14, 12, three, one, and zero citations. The total number of citations then equals 30.

- *Average number of citations per publication*. The average number of citations of the publications of a research unit. For the research unit in our example, the average number of citations per publication (30 divided by five) equals six. Without doubt, the best-known indicator based on the idea of counting the average number of citations per publication is the journal impact factor, which counts the average number of citations received by the publications in a journal. Indicators based on average citation counts are frequently used, but they are also criticised in the literature. Citation distributions tend to be highly skewed (e.g. Albarrán et al., 2011; Seglen, 1992), and therefore the average number of citations of a set of publications may be strongly influenced by one or a few highly cited publications. This is for instance observed by Aksnes and Sivertsen (2004) at the level of countries and by Waltman et al. (2012a) at the level of universities. Because of the skewness of citation distributions, suggestions are often made to replace or complement indicators based on average citation counts by alternative indicators (e.g. Aksnes & Sivertsen, 2004; Bornmann & Mutz, 2011; Leydesdorff & Opthof, 2011; Waltman et al., 2012a). Indicators based on the idea of counting highly cited publications, which we discuss below, are a frequently suggested alternative.

- *Number of highly cited publications*. The number of publications of a research unit that are considered to be highly cited, where a certain threshold needs to be chosen to determine whether a publication is counted as highly cited or not. For instance, using a threshold of 10 citations, the research unit in our example has two highly cited publications. The idea of counting highly cited publications is suggested by for instance Martin and Irvine (1983), Plomp (1990, 1994), and Tijssen et al. (2002). The i10-index reported by GS is based on the idea of counting highly cited publications.

- *Proportion of highly cited publications*. The proportion of the publications of a research unit that are considered to be highly cited. Using again a threshold of 10 citations, the proportion of highly cited publications for the research unit in our example (two divided by five) equals 0.4 (or 40%).

- *h-index*. The h-index (or Hirsch index) is defined as follows: A research unit has index *h* if *h* of its publications each have at least *h* citations and the other publications each have no more than *h* citations. For the research unit in our example, the h-index equals three. This is because the three most frequently cited publications each have at least three citations while the other two publications each have no more than three citations. The h-index was introduced in 2005 (Hirsch, 2005) and has quickly become very popular. A large number of variants and extensions of the h-index have been proposed in the literature, of which the g-index (Egghe, 2006) is probably the one that is best known. Some counterintuitive properties of the h-index are highlighted by Waltman and Van Eck (2012a). In this literature review, we do not provide a detailed discussion of the extensive literature on the h-index and its variants. Instead, see existing literature reviews (Alonso et al., 2009; Egghe, 2010; Norris & Oppenheim, 2010b; Panaretos & Malesios, 2009).

Table 1. Five basic citation impact indicators, with a distinction between size-dependent and size-independent indicators.

Size-dependent indicators	**Size-independent indicators**
Total number of citations	Average number of citations per publication
Number of highly cited publications	Proportion of highly cited publications
h-index	

In Table 1, a distinction is made between size-dependent and size-independent indicators. Size-dependent indicators aim to provide an overall performance measure. When additional publications are obtained, these indicators will never decrease. On the other hand, size-independent indicators aim to provide an average performance measure per publication. These indicators may decrease when additional publications are obtained. Size-independent indicators are typically used to make comparisons between units that are of different sizes, for instance between a small and a large research group or between a small and a large university. Most citation impact indicators for journals, such as the impact factor, are also size independent. This is because when journals are compared, we usually do not want the size of the journals (i.e., the number of publications in each journal) to have an effect on the comparison. Instead, we are usually interested in comparing journals based on their citation impact per publication. As can be seen in Table 1, the average number of citations per publication and the proportion of highly cited publications are size-independent indicators. These indicators have the total number of citations and the number of highly cited publications as their natural size-dependent counterparts. The h-index is also size dependent, but it does not have a size-independent counterpart.

1.2.2. Exclusion of specific types of publications and citations

In the calculation of citation impact indicators, often certain types of publications and citations are excluded. Below, we first discuss the exclusion of publications. We then consider the exclusion of citations.

1.2.2.1. Exclusion of publications

In the literature, various criteria for excluding certain types of publications have been proposed. Below, we discuss the most important criteria.

The most common criterion for excluding publications is based on their so-called document type. In WoS and Scopus, each publication has a document type. For instance, important document types in WoS are ‘article’, ‘review’, ‘letter’, ‘editorial material’, ‘meeting abstract’, and ‘proceedings paper’. The main reason for excluding certain document types is that publications of different document types are hard to compare with each other. This problem is of limited significance in the case of basic size-dependent indicators such as the total number of citations or the h-index, but the problem is serious in the case of size-independent indicators such as the average number of citations per publication. For instance, consider a researcher who serves as editor of a journal and who now and then writes an editorial for his/her journal. Editorials are of a very different nature from ordinary research articles, and they therefore tend to get cited much less frequently. Using a size-independent indicator such as the average number of citations per publication, a researcher would essentially be penalised for writing editorials. This can be avoided by excluding editorials from the calculation of the average number of citations per publication.

In the literature, discussions on document types and their inclusion in or exclusion from the calculation of citation impact indicators mainly relate to the WoS database. González-Albo and Bordons (2011), Zhang and Glänzel (2012), and Harzing (2013a) discuss the ‘proceedings paper’ document type. Harzing (2013a) in addition also focuses on the ‘review’ document type. The document types ‘letter’ and ‘editorial material’ are discussed by, respectively, Van Leeuwen et al. (2007) and Van Leeuwen et al. (2013). For older literature on document types in the WoS database, see Sigogneau (2000) and the references provided in this work.

Another criterion for excluding publications is the language in which publications are written. Van Leeuwen et al. (2001) and Van Raan et al. (2011) suggest that when a comparative analysis of countries or research institutions is performed, publications not written in English should be excluded from the calculation of size-independent indicators. They show that non-English language publications on average receive fewer citations than English language publications, which they suggest is because many researchers cannot read publications that are not in English. Following this reasoning, they then argue that including non-English language publications creates a bias against countries in which researchers publish a high proportion of material in their own language.

Waltman and Van Eck (2013a, 2013b) go one step further and argue that not only non-English language publications should be excluded but all publications in journals that do not have a sufficiently strong international orientation. They present criteria for identifying these journals. The possibility of excluding non-international journals is also suggested by Moed (2002) and López-Illescas et al. (2009), based on the idea that international comparisons can best be made by considering only publications in the international scientific literature. Zitt et al. (2003) reason in a somewhat similar direction. They study the effect of excluding journals with a low citation impact, which are often journals with a national focus.

1.2.2.2. Exclusion of citations

In addition to excluding certain types of publications, the suggestion is also often made to exclude certain type of citations, in particular self-citations. Self-citations can be defined at various levels, for instance at the journal level (i.e., a publication in a journal citing another publication in the same journal) or at the level of research institutions (i.e., a publication of an institution citing another publication of the same institution). However, in the literature, most attention is paid to self-citations at the level of authors. Our focus therefore is on these author self-citations.

Author self-citations are usually defined as citations for which the citing and the cited publication have at least one author name in common (e.g. Aksnes, 2003; Glänzel et al., 2004).[4] Although this is the most commonly used definition of author self-citations, some proposals for alternative definitions can be found in the literature. Costas et al. (2010) propose to distinguish between author self-citations and co-author self-citations (see also Schreiber, 2007, 2008a). From the point of view of a specific researcher, they define an author self-citation as a citation made by the researcher to his/her own work, while a co-author self-citation is defined as a citation made by a co-author of the researcher to

[4] As noted by Glänzel et al. (2004), this definition may lead to some inaccuracies. This is because a citing and a cited publication may have an author name in common, but this name may refer to two different persons who happen to have the same name. Conversely, a citing and a cited publication may not have an author name in common even though they do share an author. This may happen if an author does not write his/her name in a consistent way in different publications.

one of their co-authored works. Another proposal is made by Schubert et al. (2006), who suggest a fractional author self-citation concept based on the degree of overlap between the set of authors of a citing publication and the set of authors of a cited publication.

Regardless of the definition of author self-citations that is adopted, one needs to choose whether author self-citations should be excluded from the calculation of citation impact indicators or not. At the macro level (e.g. countries), Aksnes (2003) and Glänzel and Thijs (2004) show that the effect of author self-citations is very small. Glänzel and Thijs (2004) therefore conclude that there is no need to exclude author self-citations. Aksnes (2003) argues that below the macro level author self-citations should preferably be excluded. At the meso level (e.g. research institutions), Thijs and Glänzel (2006) are in favour of presenting citation impact indicators both including and excluding author self-citations. As an alternative to excluding author self-citations, Glänzel et al. (2006a) suggest offering supplementary indicators based on author self-citations. At the meso and micro level (e.g. individual researchers), Costas et al. (2010) consider non-self-citations to be the most relevant citations for evaluation purposes, but they emphasise that author self-citations also provide interesting information. At the micro level, Hirsch (2005) states that author self-citations should ideally be excluded, but he also claims that the h-index is not very sensitive to author self-citations, at least less sensitive than the total number of citations. Schreiber (2007) argues that Hirsch (2005) underestimates the sensitivity of the h-index to author self-citations. He prefers to exclude author self-citations from the calculation of the h-index, a position that is supported by Vinkler (2007) and Gianoli and Molina-Montenegro (2009). Schreiber (2008a) makes a similar point for the g-index, which he claims is even more sensitive to author self-citations than the h-index. On the other hand, Engqvist and Frommen (2008, 2010), Henzinger et al. (2010), and Huang and Lin (2011) suggest that the sensitivity of the h-index to author self-citations is limited and, consequently, that there may be no need to exclude author self-citations.

Fowler and Aksnes (2007) suggest that excluding author self-citations from the calculation of citation impact indicators may not be sufficient, because author self-citations may serve as an advertisement of a researcher's work and may therefore have the effect of increasing the number of citations received from others. More precisely, they indicate that each author self-citation seems to yield an additional 3.65 citations from others. Their suggestion is that there might be a need for an explicit penalty on author self-citations. An earlier study by Medoff (2006), based on a more limited data set, does not find strong evidence of an 'advertisement effect' of author self-citations.

1.2.3. Normalisation of citation impact indicators

One of the key principles of citation analysis is that citation counts of publications from different fields should not be directly compared with each other. This is because there are large differences among fields in citation density, that is, in the average number of citations per publication. For instance, a biochemistry publication with 25 citations cannot be considered to have a higher citation impact than a mathematics publication with 10 citations. There is a difference in citation density between biochemistry and mathematics of about one order of magnitude (Waltman et al., 2011b). Taking this into account, it seems we need to conclude that in our example the publication with the higher citation impact is actually the one in mathematics rather than the biochemistry one.

In addition to comparisons between publications from different fields, one should also be careful with comparisons between publications from different years. Even within the same field, a publication from 2005 with 25 citations cannot necessarily be considered to have a higher citation impact than a publication from 2010 with 10 citations. Taking into account that the publication from 2005 has had five more years to attract citations, the conclusion may be that the publication with the higher citation impact is actually the one from 2010. This would be a reasonable conclusion if we for instance know that in this field publications from 2005 on average have 40 citations while publications from 2010 on average have only five citations.

In a similar way, it is often argued that citation counts of publications of different document types, for instance the WoS document types 'article', 'letter', and 'review', should not be directly compared with each other, for instance because review articles tend to attract many more citations than ordinary research articles.

For practical purposes, there often is a need to make comparisons between publications that are from different fields or different years or that have different document types. Bibliometricians have developed normalised citation impact indicators to make such comparisons. The idea of these indicators is to correct as much as possible for the effect of variables that we do not want to influence the outcomes of a citation analysis, such as the field, the year, and the document type of a publication. Below, we summarise the literature on normalised citation impact indicators. Our focus is on normalisation for field differences. In general, normalisation for differences in publication year and document type can be performed in a similar way.

For each of the five basic citation impact indicators presented in Table 1, it is possible to develop normalised variants. We start by discussing normalised variants of the average number of citations per publication. We then consider normalised variants of the proportion of highly cited publications. Normalised variants of the size-dependent counterparts of these two indicators can be obtained in a completely analogous way (e.g. Waltman et al., 2011a) and therefore do not need any further discussion. In the context of the h-index, the third size-dependent indicator listed in Table 1, there is also some literature on the topic of normalisation (Batista et al., 2006; Iglesias & Pecharromán, 2007; Kaur et al., 2013; Radicchi et al., 2008). However, since most work on normalisation does not consider the h-index, we do not provide a further discussion of this literature.

1.2.3.1. Normalised indicators based on average citation counts

In the calculation of normalised variants of the average number of citations per publication, a key concept is the expected number of citations of a publication. The expected number of citations of a publication is defined as the average number of citations of all publications in the same field (and from the same year and of the same document type). When working with the WoS database, fields are often defined based on the WoS journal subject categories. WoS distinguishes between about 250 journal subject categories, most of which can be considered to represent a specific field of science, such as biochemistry, condensed matter physics, economics, mathematics, oncology, and sociology. Each journal covered by WoS belongs to one or more of these journal subject categories. Hence, based on the journal in which a publication has appeared, each publication indexed in WoS can be assigned to one or more journal subject categories, which then represent the field or the fields to which the publication belongs.

Given the expected number of citations of a publication, the normalised citation score of the publication is calculated as the ratio of the actual number of citations of the publication and the expected number of citations. For a set of publications of a research unit, a normalised variant of the average number of citations per publication is obtained by taking the average of the normalised citation scores of the publications of the research unit. Table 2 provides a simple example. This example considers a research unit that has five publications. For each publication, both the actual and the expected number of citations is given (first two columns of Table 2). The normalised citation score of a publication (last column of Table 2) is calculated by dividing the actual number of citations by the expected number of citations. Next, a normalised variant of the average number of citations per publication is obtained by averaging the normalised citation scores of the five publications. As shown in Table 2, the average normalised citation score equals 1.07. This score is somewhat above one, which indicates that on average the publications of the research unit have been cited above expectation.

Table 2. Example of the calculation of the average normalised citation score of a set of publications.

Actual no. of cit.	Expected no. of cit.	Norm. cit. score
14	21	0.67
12	4	3.00
3	2	1.50
1	5	0.20
0	2	0.00
Average norm. cit. score:		1.07

Another normalised variant of the average number of citations per publication is obtained by first calculating, for a given set of publications, the total number of citations actually received and the expected total number of citations and then taking the ratio of the actual and the expected total number of citations. For instance, in the case of the publications listed in Table 2, the actual total number of citations equals 30, while the expected total number of citations equals 34. Hence, the ratio of the actual and the expected total number of citations equals 30 / 34 = 0.88. The fact that the ratio is below one indicates that the total number of citations actually received is below expectation.

There is no agreement among bibliometricians regarding the question as to which of the above two normalised variants of the average number of citations per publication is to be preferred. Most bibliometricians nowadays seem to prefer the first variant, which is sometimes referred to as the average of ratios approach, over the second variant, which is sometimes called the ratio of averages approach. Using different arguments, Lundberg (2007), Opthof and Leydesdorff (2010), Van Raan et al. (2010), and Waltman et al. (2011a) claim that the average of ratios approach is more appropriate than the ratio of averages approach. However, Moed (2010b) and Vinkler (2012) present counterarguments in favour of the ratio of averages approach. Empirical comparisons between the two approaches are presented by Larivière and Gingras (2011) and Waltman et al. (2011b). They conclude that the differences between the two approaches are small, especially at the level of countries and research institutions.

In addition to the above discussion on averages of ratios versus ratios of averages, researchers have also studied various alternative approaches to calculate normalised citation scores. Lundberg (2007) suggests applying a logarithmic transformation to citation counts and normalising citation counts by calculating *z*-scores. Others have built on the work of Radicchi et al. (2008) and Radicchi and Castellano (2011), who start from the viewpoint that a proper normalisation approach should result in normalised citation distributions that are universal across fields. Radicchi et al. (2008) conclude that normalisation based on the ratio of the actual and the expected number of citations of a publication indeed yields the desired universality of citation distributions. However, Waltman et al. (2012b) claim that this conclusion is too strong and that no perfect universality of citation distributions is obtained. Abramo et al. (2012a, 2012b) compare a number of normalisation approaches and suggest that the best normalisation is obtained by dividing the actual number of citations of a publication by the average number of citations of all publications that are in the same field and that have at least one citation. Radicchi and Castellano (2012b) propose a normalisation approach that is based on a transformation of citation counts by a two-parameter power-law function. Li et al. (2013) compare this normalisation approach with a number of other approaches and conclude that, based on the criterion of universality of citation distributions, it has the best performance.

1.2.3.2. Normalised indicators based on highly cited publications

Normalised variants of the proportion of highly cited publications use a field-dependent threshold to determine whether a publication is counted as highly cited or not. The field-dependent threshold is usually chosen in such a way that the percentage of highly cited publications is the same in each field. This approach is proposed by Tijssen et al. (2002), who focus on the top 1% and the top 10% most highly cited publications in a field, and by Van Leeuwen et al. (2003), who consider the top 5% most highly cited publications. Nowadays, the idea of calculating the proportion of publications that belong to the top 10% most highly cited in their field plays an important role both in the CWTS Leiden Ranking and in the SCImago Institutions Rankings, which are the two most important bibliometric university rankings (Waltman et al., 2012a; Bornmann et al., 2012).

Choosing a citation threshold in such a way that a certain pre-specified percentage of the publications in a field, for instance 10% of the publications, is above the threshold is not entirely straightforward. It is usually not possible to obtain exactly the desired percentage of publications above the threshold. Depending on how the threshold is chosen, the percentage will be either somewhat too low or somewhat too high. The main cause of this difficulty is that there are often many publications in a field that all have the same number of citations. Because publications with the same number of citations will be either all below the threshold or all above the threshold, it becomes difficult to obtain exactly the desired percentage of publications above the threshold. There is some discussion in the literature on the best way to deal with this difficulty. Different approaches are proposed by, among others, Van Leeuwen et al. (2003), Pudovkin and Garfield (2009), Leydesdorff et al. (2011), Bornmann et al. (2012), and Waltman and Schreiber (2013). A summary of the different approaches is given by Waltman and Schreiber (2013), and an empirical comparison is presented by Schreiber (2013).

Leydesdorff et al. (2011) introduce a generalisation of the idea of identifying a certain percentage of highly cited publications in each field. Instead of making a binary distinction between publications that are highly cited and publications that are not, Leydesdorff et al. (2011) suggest defining a number of classes of publications, where each class of publications is defined in terms of percentiles of the citation distribution of a field. For instance, the first class may include all publications whose number of citations is below the 50th percentile of the citation distribution of a field, the second class may include all publications whose number of citations is between the 50th and the 75th percentile, and so on. Leydesdorff et al. (2011) propose an indicator that values publications based on the class to which they belong, with publications in the lowest class having a value of one, publications in the second-lowest class having a value of two, etc. An approach that is somewhat similar to the approach of Leydesdorff et al. (2011) is presented by Glänzel (2013). Glänzel (2013) also defines a number of classes of publications, but instead of percentiles he uses the method of characteristic scores and scales (Glänzel & Schubert, 1988) to define the classes. Publications belong to the lowest class if they have fewer citations than the average of their field, they belong to the second-lowest class if they do not belong to the lowest class and if they have fewer citations than the average of all publications that do not belong to the lowest class, and so on.

1.2.3.3. Choice of a field classification system

Normalisation of citation impact indicators, either of indicators based on average citation counts or of indicators based on highly cited publications, requires a classification system in which publications are assigned to fields. As explained above, the WoS journal subject categories are the most commonly used field classification system for normalisation purposes. However, researchers have raised some important questions related to the choice of a classification system. These questions are for instance about the sensitivity of normalised indicators to the choice of a classification system and about the possibilities for using alternative classification systems instead of the WoS journal subject categories.

Zitt et al. (2005), Adams et al. (2008), Glänzel et al. (2009), and Colliander and Ahlgren (2011) study the sensitivity of normalised indicators to the aggregation level at which fields are defined. Zitt et al. (2005) and Adams et al. (2008) observe a lack of stability of normalised indicators with respect to the aggregation level at which normalisation takes place. They argue that different aggregation levels provide different viewpoints and may all have a certain legitimacy. Glänzel et al. (2009) compare normalisation at the level of WoS journal subject categories with normalisation at higher aggregation levels defined according to the Leuven/Budapest field classification system (Glänzel and Schubert, 2003). Based on a macro level analysis of research institutions, they indicate that their preferred approach is to normalise at a relatively high aggregation level at which there are 60 fields. Colliander and Ahlgren (2011) perform an analysis of university departments and conclude that there are no substantial differences when instead of the WoS journal subject categories the 22 fields defined in the Essential Science Indicators are used for normalisation purposes.

Other analyses of the suitability of the WoS journal subject categories for normalisation purposes are reported by Van Eck et al. (2013) and Leydesdorff and Bornmann (in press). Van Eck et al. (2013) observe a strong heterogeneity in citation characteristics within medical subject categories, suggesting that the use of these subject categories for normalising citation impact indicators may be problematic. Leydesdorff and Bornmann (in press) study the way in which two fields, namely library and information science, and science and technology studies, are represented by WoS journal subject categories. They suggest that the WoS journal subject categories may be inappropriate for normalisation purposes.

Researchers have proposed various improvements of and alternatives to the use of the WoS journal subject categories for normalising citation impact indicators. Improvements are suggested by Glänzel et al. (1999) and Rons (2012). Glänzel et al. (1999) discuss the reassignment of publications in multidisciplinary journals (e.g. *Nature* and *Science*) to appropriate subject categories based on their references. Rons (2012) introduces the idea of exploiting the overlap of subject categories to obtain a more detailed classification system.

An obvious alternative to the use of the WoS journal subject categories is to replace them by an alternative field classification system. Proposals in this direction are made by Bornmann et al. (2008), Neuhaus and Daniel (2009), and Van Leeuwen and Calero-Medina (2012), who suggest the use of, respectively, Medical Subject Headings, Chemical Abstracts sections, and the EconLit classification system. An important limitation of these alternative classification systems is that each of them is restricted to a single field of science. Ruiz-Castillo and Waltman (in press) also propose the use of an alternative classification system, but instead of using an existing classification system they algorithmically construct their own classification system based on a large-scale analysis of citation relations between publications (Waltman & Van Eck, 2012b). Their algorithmically constructed classification system covers all fields of science.

A critical perspective on the normalisation of citation impact indicators is taken by Kostoff (2002) and Kostoff and Martinez (2005). They argue that the only meaningful normalisation approach is to select for each publication a small number of thematically similar publications and to compare the number of citations of a publication with the number of citations received by the selected similar publications. According to Kostoff (2002) and Kostoff and Martinez (2005), selecting similar publications needs to be done manually by experts. Colliander (in press) proposes a somewhat similar approach, but instead of selecting similar publications manually he introduces an algorithm that selects similar publications based on shared references and shared terms. The idea of comparing publications with other similar publications selected based on shared references (i.e., bibliographic coupling) is also discussed by Schubert and Braun (1993, 1996).

1.2.3.4. Alternative normalisation approaches

The normalisation approaches discussed so far are based on the idea of comparing the number of citations of a publication with the number of citations of other publications that are considered to be in the same field. We now discuss some alternative normalisation approaches that have been proposed in the literature. An attractive feature of these alternative normalisation approaches is that they do not require a field classification system.

An important alternative normalisation approach is given by the concept of citing-side normalisation. Citing-side normalisation is based on the idea that differences among fields in citation density are to a large extent caused by the fact that in some fields publications tend to have much longer reference lists than in other fields. Citing-side normalisation aims to normalise citation impact indicators by correcting for the effect of reference list length. The concept of citing-side normalisation originates from Zitt and Small (2008). Different approaches to citing-side normalisation are discussed by Zitt and Small (2008), Zitt (2010), Gómez-Sancho and Mancebón-Torrubia (2009), Moed (2010a), Leydesdorff and Opthof (2010), Leydesdorff and Bornmann (2011a), Leydesdorff et al. (2013a), Waltman et al. (2013), and Glänzel et al. (2011).[5] Empirical comparisons between citing-side normalisation and traditional cited-side normalisation are presented by Glänzel et al. (2011), Radicchi and Castellano (2012a), Leydesdorff et al. (2013b), and Waltman and Van Eck (2013a, 2013b). Radicchi and Castellano (2012a) and Leydesdorff et al. (2013b) conclude that cited-side normalisation performs better than citing-side normalisation. However, Sirtes (2012) criticises the methodology on which this conclusion is based. Waltman and Van Eck (2013a, 2013b) reach the opposite conclusion and suggest that citing-side normalisation may outperform cited-side normalisation.

Recursive citation impact indicators offer another alternative normalisation approach. These indicators give different weights to citations depending on their source. The higher the citation impact of the source of a citation, the higher the weight of the citation. As in the case of citing-side normalisation, recursive citation impact indicators correct for the effect of reference list length. The idea of recursive citation impact indicators originates from Pinski and Narin (1976). The introduction of the well-known PageRank algorithm (Brin & Page, 1998) has led to a renewed interest in recursive citation impact indicators. See Waltman and Yan (in press) and Fragkiadaki and Evangelidis (2014) for overviews of the literature on these indicators.

1.2.4. Credit allocation in the case of multi-author publications

Science is becoming increasingly collaborative. Various studies have for instance shown a continuously increasing trend in the average number of authors per publication (e.g. Gazni et al., 2012; Persson et al., 2004; Wuchty et al., 2007). Extreme examples of large-scale scientific collaboration can be found in high energy physics and in certain biomedical fields, where publications sometimes include several hundreds of authors (e.g. Cronin, 2001b).

With increasing numbers of authors per publication, it becomes more and more difficult to properly allocate the credits of a publication to the individual authors. Citation impact indicators often allocate the full credits of a publication to each individual author. This approach is known as full counting, whole counting, integer counting, or total counting. For instance, if a publication with five authors has been cited ten times, each author is considered to have ten citations. Hence, overall 50 citations are allocated to the five authors. It is clear that this approach has a certain inflationary effect, since citations received by publications with multiple authors are counted multiple times. This is sometimes

[5] Similar ideas are also suggested by Nicolaisen and Frandsen (2008), Kosmulski (2011), and Franceschini et al. (2012). However, these authors propose performing a normalisation based on the reference list length of cited publications, while citing-side normalisation is based on the reference list length of citing publications.

considered undesirable, and therefore various alternative approaches to dealing with multi-author publications have been proposed in the literature. Below, we first discuss the fractional counting approach. We then discuss approaches that take into consideration the position of an author in the author list of a publication.

1.2.4.1. Credit allocation based on fractional counting

In the fractional counting approach, the credits of a publication are fractionally allocated to the authors of the publication. Each author receives an equal share of the credits. For instance, in the case of a publication with five authors and 10 citations, each author receives one fifth of the credits of the publication, which means that each author is allocated two citations.

When working at the level of countries or institutions rather than individual researchers, there are different ways in which fractional counting can be implemented. For instance, in the case of a publication co-authored by three US researchers and one UK researcher, one possibility is to allocate the publication with weight 0.75 to the US and with weight 0.25 to the UK. Another possibility is to allocate the publication to each country with a weight of 0.5. See Gauffriau et al. (2007) for a detailed discussion of the different possibilities. Gauffriau et al. (2007) also present a systematic proposal of a terminology that can be used to distinguish between different counting approaches.

Comparisons between full and fractional counting in analyses at the level of countries are reported by, among others, Rinia et al. (1993), Gauffriau and Larsen (2005), Moed (2005a), Gauffriau et al. (2008), Huang et al. (2011), and Aksnes et al. (2012). Gauffriau et al. (2008) also provide references to earlier work in which full and fractional counting are compared. Empirical comparisons between the two counting approaches show that fractional counting yields lower citation scores than full counting. This is because publications co-authored by multiple countries on average receive more citations than publications authored by a single country. In the fractional counting approach, publications co-authored by multiple countries have less weight, and therefore fractional counting yields lower citation scores than full counting. There is no general consensus on which of the two counting approaches is to be preferred. It can be argued that full and fractional counting measure different concepts (participation vs. contribution) and both provide useful information. This perspective is emphasised by Moed (2005a). In most studies, however, a preference for fractional counting over full counting is indicated (Aksnes et al., 2012; Gauffriau & Larsen, 2005; Huang et al., 2011; Rinia et al., 1993). Full counting is often criticised because it provides non-additive statistics, with for instance the sum of the number of publications of each country in the world being larger than the total number of publications worldwide.

At the institutional level, full and fractional counting are compared by Waltman et al. (2012a) and Lin et al. (2013). In both studies, the authors express a preference for fractional counting over full counting. Waltman et al. (2012a) argue that full counting may lead to invalid comparisons across fields, even when working with normalised indicators.

At the level of individual researchers, De Solla Price (1981) argues that fractional counting is preferable over full counting. Lindsey (1980) presents an overview of bibliometric analyses at the level of individual researchers reported in the sociology of science literature. Most studies turn out to use full counting, but Lindsey (1980) argues that fractional counting is preferable. The introduction of the h-index has led to a renewed interest in counting approaches at the level of individual researchers. Fractional counting variants of the h-index are studied by Egghe (2008) and Schreiber (2008b, 2008c, 2009a). The same researchers also investigate fractional counting variants of the g-index (Egghe, 2008; Schreiber, 2009b, 2010a, 2010b).

1.2.4.2. Credit allocation based on the position in the author list

A common objection against fractional counting is that distributing the credits of a publication equally over all authors may not be fair. Some authors may have contributed more than others, and ideally this should be reflected in the way in which credit u is allocated to authors. In the literature, various approaches have been proposed for allocating credit to the authors of a publication based on their position in the author list. This is based on the idea that the position of an author in the author list of a publication provides an indication of the contribution made by that author, with the first author typically being regarded as the most important contributor. Of course, this idea is not valid in fields in which the authors of a publication tend to be ordered alphabetically. This phenomenon of alphabetical authorship is studied by Frandsen and Nicolaisen (2010) and Waltman (2012). Waltman (2012) finds that alphabetical authorship is common in mathematics, economics, and high energy physics. See Marušić et al. (2011) for a review of the literature on authorship order.

A simple approach to allocate credit to authors based on their position in the author list of a publication is to give the full credits of a publication to the first author and to give no credits at all to the other authors. This approach is known as first-author counting or straight counting. First-author counting has been studied in country-level analyses (Gauffriau et al., 2008; Huang et al., 2011; Rinia et al., 1993; Schubert et al., 1989), institutional-level analyses (Lin et al., 2013), and analyses at the level of individual researchers (Lange, 2001; Lindsey, 1980). Instead of allocating the credits of a publication to the first author, researchers have also investigated the idea of allocating the credits to the corresponding author (Huang et al., 2011; Lin et al., 2013). Another possibility is to allocate credit both to the first and to the corresponding author of a publication. Hu et al. (2010) explore this possibility in the context of the h-index. It should be noted that the concepts of first author and corresponding author can be somewhat ambiguous. Hu (2009) draws attention to the fact that an increasing number of publications have multiple first authors ('equal first authorship') or multiple corresponding authors.

Various more complex approaches to allocate credit to authors based on their position in the author list of a publication have been proposed. These approaches assign weights to the authors of a publication. The weight of an author depends on the position of the author in the author list and on the total number of authors of the publication. The typical idea is to assign the highest weight to the first author, followed by the second author, the third author, and so on. The total weight of all authors of a publication equals one, and the weight of an author determines the share of the credits of the publication that are allocated to that author. Weights can be assigned to authors in many different ways, and therefore a number of different weighted counting approaches have been introduced in the literature. These include harmonic counting (Hagen, 2008, 2010, 2013, 2014a, 2014b; Hodge & Greenberg, 1981; Jian & Xiaoli, 2013), arithmetic counting (Abbas, 2011; Egghe et al., 2000; Van Hooydonk, 1997), also known as proportional counting, geometric counting (Egghe et al., 2000), and the axiomatic counting approach of Stallings et al. (2013). Table 3 illustrates the differences between these approaches by showing the weights assigned to the authors of a publication with five authors. Other weighted counting approaches are proposed by Trueba and Guerrero (2004), Liu and Fang (2012a, 2012b), and Abramo et al. (2013).

Table 3. Weights assigned to the authors of a publication with five authors. The weights are determined based on harmonic counting, arithmetic counting, geometric counting, or the counting approach of Stallings et al. (2013).

	1st author	**2nd author**	**3rd author**	**4th author**	**5th author**
Harmonic counting	0.438	0.219	0.146	0.109	0.088
Arithmetic counting	0.333	0.267	0.200	0.133	0.067
Geometric counting	0.516	0.258	0.129	0.065	0.032
Stallings et al. (2013)	0.457	0.257	0.157	0.090	0.040

A critical perspective on weighted counting approaches is presented by Kosmulski (2012). He argues that weighted counting approaches fail to take into consideration the situation of group leaders, who in many cases are listed as the last author of a publication. When weights are assigned to authors based on their position in the author list of a publication, group leaders often will not be assigned a correct weight. Some researchers have suggested weighted counting approaches that do not depend on the order of the authors of a publication. These approaches therefore do not suffer from the group leader problem discussed by Kosmulski (2012). One approach is suggested by Tol (2011), who proposes to assign weights to the authors of a publication based on each author's past performance. Another approach, suggested by Shen and Barabási (2014), assigns weights to the authors of a publication by taking into account co-citation relations between the publication and each author's earlier work.

Discussions on weighted counting approaches often take place in the context of the h-index (Abbas, 2011; Galam, 2011; Hagen, 2008; Jian & Xiaoli, 2013; Liu & Fang, 2012a, 2012b). However, in addition to weighted counting approaches, researchers have also proposed alternative ways of correcting the h-index for the effect of co-authorship. Batista et al. (2006) and Wan et al. (2007) suggest dividing the h-index by a correction factor that depends on the number of co-authors someone has. A more complex proposal is made by Hirsch (2010), who introduces a variant of the h-index referred to as the h-bar-index. A publication contributes to someone's h-bar-index only if it also contributes to the h-bar-index of each of the co-authors of the publication.

1.2.5. Indicators of the citation impact of journals

The discussion in the previous sections has focused on citation impact indicators in general. In this section, we focus specifically on citation impact indicators for journals. We devote a separate section to this topic because of the large amount of attention it receives in the literature. See Glänzel and Moed (2002), Rousseau (2002), Bar-Ilan (2008a), and Haustein (2012) for earlier overviews of the literature on indicators of the citation impact of journals. Empirical comparisons of various citation impact indicators for journals are reported by Bollen et al. (2009), Leydesdorff (2009), and Elkins et al. (2010).

1.2.5.1. Basic citation impact indicators for journals

The best-known indicator of the citation impact of journals is the impact factor (Garfield, 1972). The impact factor of a journal equals the ratio of, on the one hand, the number of citations given in a particular year to publications in the journal in the previous two years and, on the other hand, the number of publications in the journal in the previous two years. For instance, if a journal published a total of 100 publications in 2011 and 2012 and if these publications were cited 200 times in 2013, the impact factor of the journal equals 200/100 = 2. Hence, the impact factor essentially equals the average number of citations of the publications of a journal. However, the interpretation of the impact factor as a journal's average number of citations per publication is not entirely correct. This is because in the numerator of the impact factor citations to publications of all document types are counted while in the denominator only publications of specific document types (i.e., so-called citable documents) are included (Moed & Van Leeuwen, 1995, 1996).

There is a large amount of literature on the impact factor. Here we mention only a few selected works. Garfield (1996b, 2006) discusses the history, interpretation, and proper use of the impact factor from the perspective of its inventor. More details on the history of the impact factor are provided by Bensman (2007) and Archambault and Larivière (2009). The impact factor causes a lot of debate. Some of the discussion on the impact factor is summarised by Bar-Ilan (2008a). Recently, discussion took place in a special issue of *Scientometrics* (Braun, 2012). This discussion was triggered by a critical paper about the impact factor by Vanclay (2012). It should be noted, however, that part of the debate about the impact factor is not so much about the indicator itself but more about the way in which the indicator is used for research assessment purposes. In particular, there is much criticism on the use of the impact factor for assessing individual publications (and their authors) based on the journal in which they have appeared. We will get back to this below.

In addition to the classical impact factor based on citations to publications in the previous two years, there is also a five-year impact factor, which takes into account citations to publications in the previous five years. The five-year impact factor addresses the criticism that in some fields the two-year citation window of the classical impact factor is too short (e.g. Glänzel & Schoepflin, 1995; Moed et al., 1998). See Campanario (2011) for an empirical comparison between the two-year and the five-year impact factor. The two-year and the five-year impact factor are both available in the Journal Citation Reports produced by Thomson Reuters. The Journal Citation Reports also include the immediacy index, an indicator of the frequency at which the publications in a journal are cited in the year in which they appeared. Some other citation impact indicators included in the Journal Citation Reports will be discussed below.

Various other basic citation impact indicators for journals have been proposed in the literature, either as an alternative or as a complement to the impact factor. Ingwersen et al. (2001), Frandsen and Rousseau (2005), and Ingwersen (2012) discuss a so-called diachronic variant of the impact factor. In the ordinary synchronic impact factor, citations in a single year to publications in multiple earlier years are counted. In the diachronic impact factor, citations in multiple years to publications in a single year are counted, for instance citations in 2011, 2012, and 2013 to publications in 2011. Another variant of the impact factor is introduced by Sombatsompop et al. (2004) and Rousseau (2005). They propose an ordinary synchronic impact factor, but instead of considering publications in a fixed two-year time period their proposed impact factor considers publications in a flexible journal-dependent time period. The longer it takes for the publications in a journal to be cited, the longer the time period in which publications are taken into consideration in the impact factor of the journal. In this way, the impact factor is adjusted to the specific citation characteristics of a journal. Other basic citation impact indicators for journals suggested in the literature include the share of (un)cited publications (Markpin et al., 2008; Van Leeuwen & Moed, 2005), the median number of citations (Calver & Bradley, 2009), and the h-index (Braun et al., 2006; Harzing & Van der Wal, 2009). We note that unlike most indicators for journals the h-index is size dependent. Journals with more publications tend to have higher h-indices.

1.2.5.2. Normalised citation impact indicators for journals

The citation impact indicators for journals discussed above do not correct for differences in citation density among fields. To address this limitation, a large number of normalised citation impact indicators have been proposed in the literature.

The simplest proposal is made by Pudovkin and Garfield (2004). They suggest a normalised citation impact indicator for journals that is based on the rank of a journal within its WoS subject category when journals are ordered by their impact factor. For instance, if a journal has the 10th highest impact factor within a subject category that includes 200 journals, the journal is assigned a score of (approximately) 0.95, indicating that 95% of the journals in the subject category have a lower impact factor.

Building on their earlier work (Moed et al., 1998, 1999), Van Leeuwen and Moed (2002) propose a citation impact indicator for journals that is normalised for field, publication year, and document type. Normalisation is implemented by comparing the actual number of citations of each publication in a journal with the expected number of citations, where the expected number of citations of a publication is given by the average number of citations of all publications in the same field and publication year and of the same document type. Related proposals on normalised citation impact indicators for journals are presented by Sen (1992), Marshakova-Shaikevich (1996), Sombatsompop and Markpin (2005), and Vieira and Gomes (2011). These proposals all use the WoS subject categories to define fields. Mutz and Daniel (2012a, 2012b) also suggest an approach for normalising citation impact indicators for journals. Their focus is mainly on normalisation for document type rather than normalisation for field. Leydesdorff and Bornmann (2011b) and Wagner and Leydesdorff (2012) introduce a normalised citation impact indicator for journals that, unlike most indicators for journals, is not based on average citation counts per publication. Following the ideas developed by Leydesdorff et al. (2011), the proposed indicator values the publications in a journal based on their position within the citation distribution of the field. Glänzel (2011) proposes a somewhat similar idea based on the method of characteristic scores and scales (Glänzel & Schubert, 1988).

During recent years, another approach to the normalisation of citation impact indicators for journals has been developed. This is the citing-side normalisation approach introduced by Zitt and Small (2008). The SNIP (Source Normalised Impact per Paper) indicator provided in Scopus is based on citing-side normalisation. The original version of this indicator is presented by Moed (2010a). The version that is currently included in Scopus is described by Waltman et al. (2013). See Section 1.2.3.4 for a further discussion of the literature on citing-side normalisation. Most of this literature focuses on indicators for journals.

1.2.5.3. Recursive citation impact indicators for journals

As already mentioned in Section 1.2.3.4, recursive citation impact indicators give different weights to citations depending on their source, with citations originating from a high-impact source having more weight than citations originating from a low-impact source. The idea for instance is that being cited in *Nature* or *Science* should be valued more than being cited in an obscure journal that almost no one knows about. The first proposal of a recursive citation impact indicator for journals is made by Pinski and Narin (1976). A more recent proposal, inspired by the well-known PageRank algorithm (Brin & Page, 1998), is made by Bollen et al. (2006). Recursive citation impact indicators for journals are included both in the Journal Citation Reports and in Scopus. The Journal Citation Reports include the eigenfactor and article influence indicators (Bergstrom, 2007; West et al., 2010a), while Scopus includes the SCImago Journal Rank (SJR) indicator (González-Pereira et al., 2010; Guerrero-Bote & Moya-Anegón, 2012). We now discuss these indicators in more detail. See Waltman and Yan (in press) for a more extensive overview of the literature on recursive citation impact indicators for journals.

Like the impact factor, the article influence indicator is obtained by calculating the average number of citations of the publications in a journal. However, unlike the impact factor, the article influence indicator gives more weight to citations from high-impact journals than to citations from low-impact journals. The size-dependent counterpart of the article influence indicator is referred to as the eigenfactor indicator. This indicator is proportional to the product of the number of publications of a journal and the article influence indicator. Hence, the eigenfactor indicator takes the size of a journal into account and therefore favours larger journals over smaller ones. The article influence indicator and the eigenfactor indicator have the special property that self-citations at the level of journals are not counted. Citations given by a journal to itself are ignored in the calculation of the indicators. For further discussion on the article influence indicator and the eigenfactor indicator, including empirical comparisons with other citation impact indicators for journals, see Davis (2008), West et al. (2010b), and Franceschet (2010b, 2010c, 2010d).

The SJR indicator has two versions, the original version introduced by González-Pereira et al. (2010) and the revised version discussed by Guerrero-Bote and Moya-Anegón (2012). The revised version is the one that is currently included in Scopus. The SJR indicator is fairly similar to the article influence indicator, although its mathematical definition is more complex. A special feature of the revised SJR indicator is that the weight of a citation depends not only on the citation impact of the citing journal but also on a measure of the thematic closeness of the citing and the cited journal. A citation from a citing journal that is thematically close to the cited journal is given more weight than a citation from a more distant citing journal.

1.2.5.4. Citation impact of journals vs. citation impact of individual publications

In research assessments, there is often a tendency to evaluate publications based on the citation impact of the journal in which they have appeared. Especially the impact factor is often used for this purpose. Evaluating publications based on the impact factor of the journal in which they have appeared is attractive because impact factors are easily available: more easily than statistics on the number of times individual publications have been cited. Impact factors therefore often serve as a substitute for publication-level citation statistics.

Many bibliometricians reject the use of the impact factor and other journal-level indicators for evaluating individual publications. The most important argument against this practice is that the citation impact of a journal offers only a weak predictor of the citation impact of individual publications in the journal. This is because the distribution of citations over the publications in a journal tends to be highly skewed, with for instance 20% of the publications receiving 60% of the citations. The average number of citations of the publications in a journal is therefore determined mainly by a small proportion of highly cited publications, and most publications in a journal have a citation impact that is substantially below the citation impact of the journal as a whole. Hence, the citation impact of a journal is not representative of the citation impact of a typical publication in the journal. This argument against the use of journal-level indicators for evaluating individual publications has received widespread support in the bibliometric literature. The work by Seglen (1992, 1994, 1997) on this topic has been especially influential. The inventor of the impact factor also warns against the use of this indicator for evaluating individual publications (Garfield, 1996b, 2006). Recently, the *San Francisco Declaration on Research Assessment* (http://am.ascb.org/dora/), which strongly argues against the use of the impact factor in the assessment of individual publications and their authors, received a lot of support in the scientific community.

Some researchers argue that indicators of the citation impact of journals may be useful in evaluating very recent publications. In the case of very recent publications, the number of citations received provides hardly any information, simply because there has been almost no opportunity for these publications to be cited. The citation impact of the journal in which a publication has appeared may then be seen as an interesting alternative source of information. This line of reasoning is followed by Abramo et al. (2010) and Levitt and Thelwall (2011). Abramo et al. (2010) argue that in certain fields very recent publications can better be evaluated based on the impact factor of their journal than on their individual number of citations. Levitt and Thelwall (2011) suggest evaluating recent publications using a composite indicator that takes into account both the impact factor of the journal in which a publication has appeared and the number of citations received by the publication. In line with this suggestion, Stern (2014) reports that in the prediction of the long-term number of citations of recent publications the impact factor offers useful complementary information to the short-term number of citations. On the other hand, Lozano et al. (2012) claim that since 1990 the relation between the impact factor and the number of citations of individual publications has been weakening, suggesting that the use of the impact factor as a substitute for publication-level citation statistics is becoming more and more problematic.

1.3. Evaluation practices and effects of indicator use

1.3.1. Introduction: main research strands

The use of metrics in academic evaluation and assessment systems has a range of constitutive effects (Dahler-Larsen, 2014) that have only scarcely been documented and analysed in empirical research. This knowledge gap may be due to the intimate nature of the interactions between evaluation (often confidential) and knowledge creation (often in the realm of daily life of researchers) (Wouters, 2014). In addition, the research agenda that is exploring these interactions is relatively new (Gläser et al., 2002; Gläser & Laudel, 2001; Gläser, 2010). As a result, we know some of the implications of quantitative assessments but the picture is still very incomplete (Gläser et al., 2002; de Rijcke and Rushforth, in press; Wouters 2014).

There is however a very wide-ranging set of studies that focus on the governance of science at large, within which the scarce research on effects of indicators is embedded. Providing a complete overview of this literature is not feasible because of its broad scope. Therefore, the main aim of this review is to delineate the main foci of the most important research strands.

A large body of work from higher education studies, new public management studies, organisation studies, anthropology of science, philosophy of science, economics and political science analyses effects on academic institutions of shrinking governmental research funding and the emergence of new public management from the 1980s onward. This literature characterises the rise of performance measurement in academic settings as part of a broader surge of accountability measures that has swept across public institutions over the past three decades (cf. Feller, 2009; Keevers et al., 2012; Krücken et al., 2013; Mirowski, 2011; Nedeva et al., 2012; Radder, 2010; Schimank, 2005) – leading some commentators to claim that we now inhabit 'audit' and 'evaluation' societies (e.g. Power, 1997; Dahler-Larsen 2012; Strathern, 2000). The most critical voices portray academics as becoming dominated (albeit sometimes willingly) by pre-defined, measurable outcomes that fulfil informational needs of a neo-liberal higher education system (Craig et al., 2014; Sauder & Espeland, 2009; Shore, 2010; Burrows, 2012). As such, indicators are positioned as tools to steer academic institutions and researchers towards becoming more like market-oriented actors, by actively stimulating competition, instrumentality and privatisation strategies (cf. Deem, Hillyard and Reed 2007; Willmott, 2011; Leisyte & Dee, 2012).

Science policy and higher education studies (e.g. De Boer et al., 2007; Rebora & Turri, 2013; Reale & Seeber, 2013) typically focus on formal characteristics of the assessment mechanisms adopted in national or regional evaluation systems (Fealing, 2011; Reale & Seeber, 2013). Most research in this area provides the impetus for furthering science policy, and puts much effort into methods of research evaluation. These studies often respond relatively swiftly towards the introduction of new evaluation programmes and methods (cf. Cozzens & Melkers, 1997). A typical example is the study by Luukkonen (2012), that examines the peer review processes of the then recently established European Research Council (ERC). Science policy studies tend to focus mainly on formalised national evaluation systems, in which indicators of funding success play a larger role than citation based indicators. Even in indicator based systems such as the Australian evaluation system citation analysis does not figure prominently (Gläser et al., 2010). This does not mean that indicators are not used at all, but they are often not visible in the formal systems of accounting (Wouters 2014, p. 49-50).

Sociology of science literature, on the other hand, focuses on institutional and organisational dynamics of science and innovation, including managerial and control mechanisms (Whitley & Gläser, 2007; Whitley, 1984, 2000, 2011; Whitley et al., 2010). Organisational approaches to the sociology of science teach us that scientific quality control is a thoroughly social and organisational phenomenon, and not exclusively cognitive/epistemological (cf. Hemlin, 2006; Whitley, 2000). This work intimates knowledge production is changing in light of transformations to relational systems in which academic researchers are doing their research (Whitley et al., 2010; Musselin, 2013; Paradeise & Thoenig, 2013). It has also begun to analyse how researchers handle the demand for accountability in their epistemic decision-making. However, most of these studies do not deal directly with effects of indicator uses. The studies that do engage with indicator effects will be discussed in the next section.

1.3.2. Known effects

This section starts with an overview of what is known about the effects of evaluative metrics in research practices (including strategic behaviour and goal displacement, task reduction, and potential biases towards interdisciplinary research). This is followed by a discussion of empirical research on institutional responses to metrics-based assessments. Lastly, emerging research on the complex relation between indicators and knowledge production is described.

1.3.2.1. Strategic behaviour and goal displacement

In tandem with the development and first applications of performance indicators in the 1970s, discussions surfaced about strategic behaviour and ‘gaming the system’ (cf. MacRoberts & MacRoberts, 1989). Researchers are not passive recipients of research evaluation but play an active role in assessment contexts (cf. Aksnes & Rip, 2009; Van Noorden, 2010). Assessment systems that affect money or reputation, whether the systems are peer review or indicator based, will tend to influence researchers’ behaviour in two ways (Butler, 2007). The first is goal displacement: scoring high on the assessment criteria becomes the goal rather than a means of evaluating if certain objectives (or performance levels) have been met (Colwell et al., 2012, p. 27). The second is a more fundamental transformation of the scientific or scholarly process itself in response to the assessment criteria (for instance by avoiding risk in selecting research topics), a transformation that may be harder to recognise (Butler, 2007, p. 572).

Studies that focused on effects of funding and evaluation regimes on scientific output have indeed demonstrated goal displacement. Butler (2003; 2005) for instance analysed the introduction of performance metrics in Australian research funding allocation. Her study revealed a sharp rise in publications tracked by bibliographic databases in all university fields (but not in other branches of research where this type of funding allocation is not present) when funding becomes linked with publications (Butler, 2005). As in the Australian evaluation system there was no differentiation between publications (besides being peer-reviewed or not), the amount of publications especially rose in the third quartile of journals that are more easily accessible. Butler earlier demonstrated how this strategy, while leading to a rise of the relative share of Australian publications, has also contributed to a decline of scientific impact (measured in citations) during the same period (Butler, 2003). Colwell et al. (2012) state that researchers' quality considerations may be displaced by incentives to produce higher quantities of publications when funding is explicitly linked to research output (in terms of the number of publications) (Colwell et al., 2012, p. 26). Similar effects of the use of bibliometrics on the amount of publications have been found in Spain (Jiménez–Contreras et al., 2003), Sweden (Hammarfelt & De Rijcke, 2015), Denmark (Ingwersen & Larsen, 2014), Flanders, and Norway (Aagaard et al., 2015; Ossenblok et al., 2012; Schneider, 2009).

Strategic response by the research community was also demonstrated in a longitudinal bibliometric study of UK science covering almost twenty years (Moed, 2008). UK publication patterns between 1985 and 2003 suggests that specific publication patterns emerged in years before three RAEs that took place in that period (1992, 1996, 2001), depending on whether the RAE was aimed at quantity or quality of publications. In the UK, findings of 'playing the RAE game' (Harley, 2002) in this way are numerous (e.g. Hare, 2003; Keenoy, 2005; Alldred & Miller, 2007; Sousa & Brennan, 2014). Another study of the UK RAE impact shows that the cumulative research productivity of individuals increased over time, but the effects differed across departments and individuals. Where researchers in higher-ranked programmes increased their output in higher-quality journals, researchers in lower-ranked departments aimed at increasing their publications in other outlets (Moore et al., 2002). A survey among journal editors conducted at the end of the 1990s (Georghiou et al. 2000) also showed that the RAE influenced where authors published. Indeed, research shows that the status of a journal is crucial for academics' submission decisions (Harley et al., 2010; Chew et al. 2007). However, claims that the RAE would lead towards salami publishing of 'least publishable units' (cf. Huth, 1986) were not confirmed in a later study (Georghiou et al., 2000).

1.3.2.2. Biases against interdisciplinarity

A related concern is the potential influence of disciplinary assessments like the REF on interdisciplinary research. An early survey of the impact of the 1996 UK RAE reported evidence of negative effects for interdisciplinary work. Almost half of those in management positions felt the RAE 'had hindered' interdisciplinarity (Mcnay, 1998, p. 20). A worldwide survey among demographers showed contradictory results. It revealed that this community displayed no tendency to focus on monodisciplinary research in terms of reading or publishing activity (Van Dalen & Henkens, 2012). In economics and many departments in business studies, however, publication productivity has been strongly stimulated by the ubiquitous use of journal rankings as obligatory publication outlets for faculty. These lists are not based on citations but on a qualitative consensus in mainstream economics and business studies about the top journals (more recent lists use a version of the Journal Impact Factor). In a first comparative analysis of the effect of these rankings in business and innovation studies, they were found to be biased against interdisciplinary work (Rafols et al., 2012). This study concludes that citation indicators may be more suitable than peer review for interdisciplinary work because criteria of excellence are essentially based on disciplinary standards.

Pontille and Torny (2010) analysed the production of three journal lists in the humanities and social sciences by the Australian Research Council (ARC), the European Science Foundation (ESF), and the French Agency for Evaluation Research and Higher Education (AERES). They found that the production of journal ratings is a highly cognitive and political task, and not merely a matter of univocal inventory-making. The modalities that were selected produced very different effects in terms of which scientific communities were involved, how boundaries were drawn around disciplines, and the ways in which revision processes adopted criticism from the fields involved. According to the authors, these and other built-in tensions will not only keep feeding critique but also the need for permanent revisions of the ratings (see also Jensen, 2011).

As soon as research managers started to set publication targets based on the two top categories on the list, ARC decided to drop the established journal rankings in its assessment system (Colwell et al., 2012, p. 56).

1.3.2.3. Task reduction

An analysis of researchers' responses to funding criteria in Australia has analysed the extent to which researchers are forced to focus their tasks and types of publication (Laudel & Gläser, 2006). A tension was observed between the formalised journal ranking procedures and researchers' own evaluations of their work. In 10 out of the 21 disciplines under study, the four publication types used in the Australian evaluation were not the same as the four types of output that researchers themselves found most important. This mismatch may result in the abandonment of particular types of work (Laudel & Gläser, 2006, p. 294). Laudel and Gläser conclude that arts and humanities are likely to suffer greatly in situations whereby evaluation systems have been modelled primarily around journal oriented disciplines.

One could feasibly expect that increased drives towards journal publications will devalue other academic activities. A worldwide survey among demographers in developed and developing countries (also discussed in the previous section) found that ‘traditional’ academic tasks – such as writing referee reports or translating research outcomes for policy – are negatively affected by a move toward rewarding individual productivity (Van Dalen & Henkens, 2012). A study of the RAE exercise in 1996 found for instance that publication in professional journals was ‘actively discouraged’ by some research managers (Mcnay, 1998, p. 22). Hoecht (2006) argues that the increased audit-based quality control has replaced a trust-based form of control and as such autonomy has decreased in academic settings in the UK in recent decades. This form of one-way accountability that lacks a fostering of trust, Hoecht warns, might negatively affect innovative teaching and research practices. Similarly, Willmott (2011) argues that audit-based control pushes academics towards mainstream topics that have the highest chance of being published in the highest ranked journals.

An exploratory study of UK university education departments shows that the use of performance indicators in evaluation is perceived as an increasing recognition of specific forms of academic involvement – academic research and publications – that are often opposed to more public roles, including applied research, writing professional publications, or teaching (Wilson & Holligan, 2013). The authors argue that the ways in which this development is perceived depends on the position of the academics in question and whether or not it fits with their scholarly practices.

In a recent study Laudel and Gläser (2014) find that the chances of ‘exceptional’ projects (in terms of planned innovations and answers to ‘big questions’) being funded appears to reduce across all disciplines in an increasingly standardised grant funding landscape, even in those funding programmes that enable this type of research (Laudel & Gläser 2014, p. 1208).

1.3.2.4. Effects on institutions

In addition to possible goal displacement, bias against interdisciplinarity and reduction of task complexity, metrics-informed assessments have also been thought to affect institutional arrangements as well as the relationship between higher education and government (Martin & Whitley, 2010). According to Colwell et al. (2012), the RAE has unintentionally created a ‘transfer market’ in faculty in the UK (see also Elton, 2000). The RAE has led universities to develop strategies regarding hires, some focusing on recruiting younger staff with research potential, with others hiring only ‘well-established researchers’ in the run-up to the RAE (Colwell et al., 2012, p. 27/28). This second strategy was also observed in Australian universities’ responses towards formula-based funding (Gläser et al., 2002, p. 17). This type of strategic behaviour may have long-lasting effects on the position of universities and research institutes and thereby on the research agenda (Wouters, 2014).

Alignment between demands of various formal evaluation agencies along more or less the same standardised criteria is likely to focus the attention of organisations and sub-groups towards satisfying them (Pfeffer & Salancik, 2003). Seeing that epistemic properties of knowledge and research funding and organisation are strongly connected (Laudel & Gläser, 2014), there is reason to assume that the development of formal performance criteria has grown stepwise with local lock-in mechanisms dependent on specific systems of governance, rather than based on a consistent evaluation logic (cf. Van der Meulen, 2007). Although the rise of formal evaluation may appear an isomorphic phenomenon, the extent to which it has transformed scientific institutions is far from uniform (Whitley & Gläser, 2007). So far little is known of the capacities that governance mechanisms like evaluation programmes have for actually controlling and steering loosely coupled work systems of academic knowledge (Gläser, 2013).

Knowledge about the performativity of numbers suggests that the availability of metrics for generating and ordering hierarchically information about performance creates a demand for such information (Porter, 1995; Wouters, 2014). Such information-generating functions could carry authority even if some of its first-order epistemic limits are known (Dahler-Larsen, 2012). Organisations cannot resist the temptation to collect such information because it is considered as strategically useful in managing researchers and improving on measures of organisational performance. The fact that competitors collect similar information for strategic purposes makes the prospect of opting-out or ignoring such information perilous (see also effects of rankings e.g. Espeland & Sauder, 2007). The legitimacy of such indicators does not rest exclusively on their first-order accuracy, but also on the fact that they are assumed to carry authority within the institutional environments with which organisations strategically engage.

Conversely, emerging empirical work on rankings shows how in complex academic settings performance metrics are not tightly coupled with actions across all sections of the organisation (cf. De Rijcke et al., 2015). This goes against recent tendencies to assume that patterns of behaviour towards indicators at one level of a public organisation imply uniformity across the board (e.g. Sauder & Espeland, 2009).

More fundamental changes in the character of academic settings as independent and critical institutions (Shore, 2008, 2010; Craig et al., 2014) as well as increased levels of stress and anxiety are also reported in personal accounts and interview-based analyses (Chandler et al., 2002; Gill, 2009; Sá et al., 2013). Burrows (2012) argues that performance-based control mechanisms have become autonomous entities that are increasingly used outside the original context of evaluations, and get a much more active role in shaping the everyday work of academics. According to Burrows (2012), the rise of numbers into the fabric of university bureaucracies may even create feelings of powerlessness among academics. Neoliberal universities provide fertile ground for a 'co-construction of statistical metrics and social practices within the academy.' (Burrows, 2012, p. 361) Among other things, Burrows contends that this leads to a reification of individual performance measures such as the h-index (ibid, p. 361). Though hard to assess, such statements of discontent expressed by researchers are empirically important and are therefore included in this review. The emerging empirical evidence discussed in the final section below suggests that the use of metrics in decision-making contexts cannot however be explained simply as explicit responses to top-down commands. To some extent metrics also seem to transform more fundamentally 'what can be talked about' and how valuations were being made (cf. Espeland & Stevens, 1998; Lamont, 2012).

1.3.2.5. Effects on knowledge production

The evidence discussed above indicates that performance-informed research assessment does indeed increase the pressure on researchers and institutions to meet the performance criteria, irrespective of whether the latter are based on peer review or on citations. That the research community and other stakeholders respond strategically towards interventions deliberately designed to align them with current science and innovation policy priorities (Whitley & Gläser, 2007) may in turn have unintended effects. Examples include the mechanisms of goal displacement and the more structural changes to research priorities, publication activities, and research capacity building and organisation. However, the evidence to support these claims is at best partial. Whilst most existing studies focus on systems where funding decisions are based directly on performance, whether this link is necessary for unintended effects to occur is debatable. The effects are not primarily based on the amount of funding that is shifted due to performance differences, but on the effect it has on researchers' reputation (Hicks, 2012). Systems where performance is publicly reported but not directly linked to funding may therefore lead to comparable or identical effects. Lastly, the most visible types of strategic behaviour may obscure more fundamental shifts in knowledge production (Wouters, 2014).

Butler (2003, 2005) notes conservatism of metrics users as a long-standing problem, in displaying a preference for user-friendly measurements that trump other inclinations for adopting more state-of-the art scientometrics. Many scientometric contributions take a normative stance regarding 'unanticipated effects' of quantitative performance measures on the scientific system and debate whether the field itself might play a more active role in promoting 'good practices' (Van Dalen & Henkens, 2012; Garfield, 1996a, 2006; Weingart, 2005). This concern can be seen in one of the leading journals – *Scientometrics* – that recently published an issue dedicated solely to the uses and misuses of the Journal Impact Factor (JIF) (Braun, 2012). Many studies have highlighted epistemic limitations in the JIF, particularly in the evaluation of individuals (cf. Moed and Van Leeuwen, 1996; Buela-Casal and Zych, 2012; Seglen, 1992, 1994, 1997; Simons, 2008), whilst others cite ‘gaming’ by journal editors and publishers as a perverse effect of its rise to prominence in academic evaluation contexts (Archambault & Larivière 2009, p.635).

However, emerging empirical evidence paints a more complex picture of how certain metrics, including the JIF, become reified in research management and decision-making contexts, as both formal and informal standards against which to assess the value and usefulness of research activities (e.g. Aksnes & Rip, 2009; Buela-Casal & Zych, 2012; Derrick & Gillespie, 2013; Rushforth & De Rijcke, n.d.; Sá et al., 2013; Stephan, 2012).

At least two potential explanations are provided in these literatures for the reification of evaluative indicators (both ‘amateur’ and ‘advanced’). First of all, the metrics inform deep-seated, firmly established mechanisms to build reputation and to hire, select and promote staff (including publishing in high Impact Factor, peer reviewed journals). Secondly, the responsibility for certain applications of bibliometric indicators is spread over many key stakeholders in the current ‘citation infrastructure’ (Wouters, 2014) including scientometricians, publishers, librarians (cf. Åström and Hansson 2013; Demšar and Južnič 2013; Petersohn 2014), policy makers, evaluators, research managers, consultancies, researchers, and other metrics users (De Rijcke & Rushforth, in press). Changing the current dominant ‘order of worth’ in research assessment (cf. Boltanski & Thévenot, 2006; Linkova, 2014; Stöckelová, 2014) is therefore also a distributed responsibility.

An exploratory study in Dutch biomedicine (Rushforth & De Rijcke, n.d.) details some of the conditions under which performance metrics are more or less likely to become routinised or peripheral to knowledge production processes (Colyvas, 2012). The authors found that prestige – an outcome of knowledge work which gets recycled as an ‘exchange good’ by academic scientists (Stephan, 2012) – was tightly coupled with citation counts and indicators like the JIF and the h-index (in targeting specific publication outlets, referencing ‘hot’ papers, negotiation of authorship priority etc.). Quantitative indicators were also observed *less formally* to feed into quite routine knowledge-producing activities on the ‘shop-floor’ (e.g. discussions over whom to collaborate with and when, how much – additional – time to spend in the laboratory producing data). The JIF in particular functioned on occasion as a screening device for selecting useful information from the overwhelming amounts of literature scientists could potentially read. These examples suggest that metrics-criteria feed into deliberations over where new scientific knowledge is likely to emerge or can be found.

Generally speaking, the promise of bibliometric tools for ‘reducing complexity’ is a feature that science policymakers and managers find most appealing (Cronin & Sugimoto, 2014; Woelert, 2013). Journal ranking tools like the JIF help make ‘commensurable’ the levels of prestige acquired from publishing in one journal over another (Espeland & Stevens, 1998). This ‘shortcut’ is particularly attractive where there is a lack of substantive expertise in a field or of informal knowledge of the reputational standings of journals among peers.

Empirical research does however reveal a discrepancy between the importance of indicators in evaluation practices according to academics and their own judgment of the accuracy of certain measures (Buela-Casal & Zych, 2012). For instance, the JIF can both denote a certain standing in a field and a particular type of scientific work (e.g. descriptive versus causal) (Rushforth & De Rijcke, n.d.). In addition, some researchers seem willing to wait longer for editorial decisions when the JIF is higher, which shows the clogging effects JIFs can have on scholarly communication (Rousseau & Rousseau, 2012). Furthermore, research suggests that scientists often have ambivalent attitudes about performance and citation indicators (Hargens and Schuman, 1990). Researchers often preserve prior beliefs about the value of their work in interpreting citation scores. Citations acquire value as part of the reward system and are often mobilised by scientists in pursuit of scarce resources, while at the same time they are criticised for not reflecting actual scientific contribution (Aksnes & Rip, 2009, p. 895). This is explained in the literature by the concept of 'folk citation theories', which do not have to be consistent in order to be mobilised by researchers as explanatory devices in their competition for reputation. However, the sophistication and complexity of scientists' interpretation of citation should not be underestimated. Aksnes & Rip (2009) observe that their respondents know the complexity and ambiguity in citations. 'In other words, scientists have a sophisticated understanding of the citation process and its outcomes, and can explicate such understanding when there are no immediate stakes to be defended.' (Aksnes & Rip, 2009, p. 904).

Lastly, the use of performance indicators and more advanced forms of bibliometric information may also influence the terms and conditions of the development of research agendas. Derrick and Pavone (2013) have reviewed government policies in three countries (the UK, Australia, and Spain). Each country has committed itself to increasing the societal role of researchers by opening up the setting of research problems by a larger group of stakeholders. The study characterises this as a move towards more participatory approaches. The authors argue that the use of scientometrics for informing this process has so far been largely overlooked. The question is too often formulated as the choice between peer review and metrics (Taylor, 2011). Derrick and Pavone (2013, p. 573) conclude that the choice should be about the type of use that scientometric information is put to. 'The simple information portrayed by correctly calculated and applied bibliometric indicators has the potential to engage a larger group of stakeholders than previous evaluation systems could.' According to this study, future research policies should take advantage of bibliometrics to foster greater democratisation of research to create more socially reflexive evaluation systems.

1.4. Bibliometrics and the use of indicators summary

1.4.1. Bibliographic databases

The three most important multidisciplinary bibliographic databases are WoS), Scopus, and GS. Scopus has a broader coverage of the scientific literature than WoS. Almost all literature indexed in WoS is also covered by Scopus, but the reverse is not the case. Some studies report that journals covered by Scopus but not by WoS tend to have a low citation impact and tend to be more nationally oriented, suggesting that the most important international scientific journals are usually covered by both databases. GS is generally found to outperform both WoS and Scopus in terms of its coverage of the scientific literature. However, there are a few fields, mainly in the natural sciences, in which some studies report the coverage of GS to be worse than the coverage of WoS and Scopus. On the other hand, the coverage of GS has been improving over time, so it is not clear whether the limited coverage of GS in natural science fields still applies today. GS is often criticised for its lack of quality control and transparency. GS also has the disadvantage of being difficult to use for large-scale bibliometric analyses.

SSH create special challenges for bibliometric analyses. Book publications and publications in national journals play an important role in SSH. These publications are often not indexed in bibliographic databases. Bibliometric analyses in computer science and engineering involve similar difficulties. Many computer science and engineering publications appear in conference proceedings. Proceedings literature tends to be less well covered by bibliographic databases, especially by WoS and Scopus, than journal literature. Another problem related to proceedings literature is that the same work may be published multiple times, for instance first in a conference proceedings and then in a journal.

1.4.2. Basic citation impact indicators

A large number of citation impact indicators have been proposed in the literature. Most of these indicators can be seen as variants or extensions of a limited set of basic indicators. These basic citation impact indicators are the total and the average number of citations of the publications of a research unit (e.g. of an individual researcher, a research group, or a research institution), the number and the proportion of highly cited publications of a research unit, and a research unit's h-index. There is criticism in the literature on the use of indicators based on total or average citation counts. Citation distributions tend to be highly skewed, and therefore the total or the average number of citations of a set of publications may be strongly influenced by one or a few highly cited publications ('outliers'). This is often considered undesirable. Indicators based on the idea of counting highly cited publications are suggested as a more robust alternative to indicators based on total or average citation counts. The h-index offers another alternative.

1.4.3. Exclusion of specific types of publications and citations

In bibliometric analyses, it needs to be decided which types of publications and citations are included and which are not. In WoS and Scopus, each publication has a document type. It is clear that research articles, which simply have the document type 'article', should be included in bibliometric analyses. However, publications of other document types, such as 'editorial material', 'letter', and 'review', involve more difficulties and may be either included or excluded. It is also suggested in the literature that in some bibliometric analyses non-English language publications or publications in national journals should not be included. In international comparisons, including these publications may create a bias against countries that have many of these publications.

Most bibliometric researchers prefer to exclude author self-citations from bibliometric analyses. There is no full agreement in the literature on the importance of excluding these citations. In some bibliometric analyses, the effect of author self-citations is very small, suggesting that there is no need to exclude these citations. In general, however, it is suggested that author self-citations should preferably be excluded, at least in analyses at low aggregation levels, for instance at the level of individual researchers.

1.4.4. Normalisation of citation impact indicators

In research assessment contexts, there is a frequent need to make comparisons between publications from different scientific fields. There is agreement in the literature that citation counts of publications from different fields should not be directly compared with each other. This is because there are large differences among fields in the average number of citations per publication. Researchers have proposed various approaches to normalise citation impact indicators for field differences. In addition to normalising for field differences, these approaches often also normalise for differences between older and more recent publications and for differences between publications of different document types, for instance research articles and review articles.

Most attention in the literature has been paid to normalised indicators based on average citation counts. Recent discussions focus on various technical issues in the calculation of these indicators. As mentioned above, indicators based on highly cited publications are sometimes considered preferable over indicators based on average citation counts. Normalised indicators based on highly cited publications for instance count the proportion of the publications of a research unit that belong to the top 10% or the top 1% of their field. There are technical discussions in the literature on the best way to calculate these normalised indicators. Researchers also study more sophisticated variants of these indicators. In these variants, publications can be valued in a flexible way based on their position within the citation distribution of their field.

A key issue in the calculation of normalised citation impact indicators is the way in which the concept of a scientific field is operationalised. The most common approach is to work with the predefined fields in a database such as WoS, but this approach is criticised a lot. Some researchers argue that fields may be defined at different levels of aggregation and that each aggregation level offers a different but legitimate viewpoint on the citation impact of publications. Other researchers suggest the use of disciplinary classification systems (e.g. Medical Subject Headings or Chemical Abstracts sections) or sophisticated computer algorithms to define fields, typically at a relatively low level of aggregation. Yet another approach is to calculate normalised citation impact indicators without defining fields in an explicit way. This idea is implemented in so-called citing-side normalisation approaches, which represent a recent development in the literature.

1.4.5. Credit allocation in the case of multi-author publications

The average number of authors of publications in the scientific literature keeps increasing, indicating a trend toward more and more collaboration in science. This trend makes it increasingly difficult to properly allocate the credits of a publication to the individual authors. The most common approach is to allocate the full credits of a publication to each individual author. This approach is known as full counting. In this approach, the citations to a multi-author publication are counted multiple times, once for each of the authors, even for authors who have made only a small contribution. Because the same citations are counted more than once, the full counting approach has a certain inflationary effect, which is sometimes considered undesirable. A number of alternative credit allocation approaches have therefore been proposed in the literature.

In the fractional counting approach, the credits of a publication are shared equally by all authors. For instance, in the case of a publication with five authors and 10 citations, each author is allocated two citations. A number of studies reported in the literature express a preference for fractional counting over full counting, but there is no general consensus on this issue. Another approach suggested in the literature is to fully allocate the credits of a publication to the first author. This is based on the idea that the first author of a publication is the most important contributor. In fields such as mathematics, economics, and high energy physics, in which the authors of a publication are often ordered alphabetically, this idea of course is not applicable. An alternative possibility is to fully allocate the credits of a publication to the corresponding author instead of the first author. A final approach discussed in the literature is to allocate the credits of a publication to the individual authors in a weighted manner, with the first author receiving the largest share of the credits, the second author receiving the second-largest share, and so on.

1.4.6. Indicators of the citation impact of journals

The best-known indicator of the citation impact of journals is the impact factor. There is a lot of debate about the impact factor, both regarding the way in which it is calculated and regarding the way in which it is used in research assessment contexts.

Various improvements of and alternatives to the impact factor have been proposed in the literature. It is for instance suggested to take into account citations during a longer time period, possibly adjusted to the specific citation characteristics of a journal, or it is proposed to consider the median instead of the average number of citations of the publications in a journal. Another suggestion is to calculate an h-index for journals as an alternative or complement to the impact factor.

Researchers also argue that citation impact indicators for journals need to be normalised for differences in citation characteristics among fields. A number of normalisation approaches have been suggested. One of these approaches is implemented in the SNIP indicator available in Scopus.

Another idea proposed in the literature is that in the calculation of citation impact indicators for journals more weight should be given to citations from high-impact sources, such as citations from *Nature* and *Science*, than to citations from low-impact sources, for instance citations from a relatively unknown national journal that receives hardly any citations itself. This idea is implemented in the eigenfactor and article influence indicators reported, along with the impact factor, in the Journal Citation Reports. The same idea is also used in the SJR indicator included in Scopus.

The impact factor and other citation impact indicators for journals are often used not only in the assessment of journals as a whole but also in the assessment of individual publications in a journal. Journal-level indicators then serve as a substitute for publication-level citation statistics. The use of journal-level indicators for assessing individual publications is rejected by many bibliometricians. It is argued that the distribution of citations over the publications in a journal is highly skewed, which means that the impact factor and other journal-level indicators are not representative of the citation impact of a typical publication in a journal. Some bibliometricians agree with the use of journal-level indicators in the assessment of very recent publications. In the case of these publications, citation statistics at the level of the publication itself provide hardly any information.

1.4.7. Main research strands on indicator effects

The constitutive effects of performance metrics have only scarcely been documented and analysed in empirical research. There is a wide-ranging set of literatures that focus on the governance of science at large. A multi-disciplinary body of work (e.g. from new public management studies, organisation studies, anthropology of science, political science) portrays the rise of performance measurement in academic settings as part of a broader upsurge of accountability measures in public institutions from the 1980s onward. Indicators are positioned as tools that drive competition, instrumentality and privatisation strategies and help steer academic institutions and researchers towards becoming more like market-oriented actors. Science policy studies focus mainly on existing and new types of formal assessment tools, and on providing methods for research evaluation (Fealing et al. 2011). They concentrate on formalised national evaluation systems, in which citation-based indicators do not play a prominent role. Sociology of science intimates scientific quality control as a thoroughly social and organisational process. The field analyses accountability measures in light of transformations to institutional and organisational dynamics of science and innovation. Most studies do not deal concretely with effects of indicator uses.

1.4.8. Strategic behaviour and goal displacement

Studies about the effects of funding and evaluation regimes on research output confirm the presence of goal displacement in a number of countries (a process in which scoring high on performance measures becomes a goal in itself, rather than a means of measuring whether a certain performance level has been attained). Many studies of UK research assessment systems and output patterns provide evidence of strategic response by the research community in terms of patterns of output production. There are also indications that the RAE affected researchers' choices for particular outlets, but salami slicing was not confirmed.

1.4.9. Effects on interdisciplinarity

Evidence for the impact of disciplinary assessments like the REF on interdisciplinary research varies. Studies show that negative effects such as goal displacement do occur, but the dynamics are most likely discipline-specific. In business and innovation studies, for instance, the use of journal ranking lists was found to be biased against interdisciplinary work. These lists are not based on citations but on a qualitative consensus in a field.

1.4.10. Task reduction

Studies confirm the abandonment of particular types of work (e.g. teaching, outreach), and a focus on particular publication forms (international, peer-reviewed journal articles) and types of research topics (the mainstream). Again, the known effects vary per discipline and the evidence is limited.

1.4.11. Effects on institutions

The rise of formal evaluation may appear isomorphic, but existing research does not reveal uniformity in institutional-level transformations of governance structures. Some studies do find an alignment between institutional measures and the standardised criteria of formal evaluation agencies (e.g. mirroring institutional hiring or promotion strategies with criteria set by funders). Some analysts explain that the legitimacy of indicators in generating and ordering strategic information on performance rests for a large part on the authority they carry within institutional environments –

relatively independent of any first-order (in) accuracies. The functioning of bibliometric tools in 'reducing complexity' is frequently cited as a reason for their widespread appeal among policymakers and research managers.

1.4.12. Effects on knowledge production

Emerging empirical evidence suggests that the use of metrics in decision-making contexts cannot simply be explained as explicit responses to top-down commands. Recent studies find a reification of evaluative metrics in research management and decision-making contexts, as formal and informal standards against which research activities are assessed. These studies suggest that in some fields, performance metrics have become routinised elements of relatively mundane knowledge production processes. Other incipient analyses point to a discrepancy between the importance of indicators in evaluation practices according to academics and their own judgment of the accuracy of certain measures. The sophistication of academics' understanding of the (dis-)advantages of performance measures should however not be underestimated. Lastly, the use of performance indicators and advanced bibliometric information may also influence the conditions under which research agendas are developed (e.g. scientometric information may potentially play a role in democratising the process of agenda-setting). Again, the evidence is fragmented and not complete.

2. Peer review and bibliometrics

2.1. Forms of peer review

Peer review is without doubt the most important method of quality control in the sciences, the social sciences, arts and the humanities (Bornmann, 2011a; Daston & Galison, 2007; Holbrook, 2010; Lee et al., 2013). Although it is often referred to as if it can be interpreted as a standardised procedure of quality control – in the same way as 'the scientific method' is cited as a guarantee of truthfulness – in fact the concept of peer review has a variety of meanings. Peer review refers to a variety of scholarly and scientific practices that can display huge differences in their scientific rationale, their quality criteria, their social organisation, and their relationship to the scientific and scholarly work that is being evaluated (Chubin & Hackett, 1990). Since the emergence of the history and sociology of science in the 1930s (Merton, 1973), and the interdisciplinary field of science and technology studies in the 1960s to 1980s (Knorr-Cetina & Mulkay, 1983; Latour & Woolgar, 1979), processes of peer review have been studied in more detail from a variety of perspectives. Moreover, the increasing scale and complexity of the international research endeavour has diversified its goals and organisation and has made it more urgent to guarantee the quality of quality control. As a result, more stakeholders have an interest in the study of peer review and it has created a huge and diverse literature on the topic.

First of all, peer review has been classified according to the objectives of the review (Geisler, 2000; Wager et al., 2002):

1. assess the quality of research results, outcomes, projects and programmes;
2. determine the level of performance, either in absolute terms or comparatively, of (parts of) the scientific and innovation system;
3. promote accountability;
4. contribute criteria and evidence for resource allocation;
5. contribute criteria and evidence for science and technology policy making;
6. contribute criteria and evidence for career decisions and human resource policies.

Peer review has also been classified according to the moment it takes place in the cycle of scientific knowledge creation (Moed, 2005a):

1. peer review of grant proposals in the context of funding decisions;
2. peer review of manuscripts in the context of publication decisions by journal or book publishers;
3. peer review of scientific data in the context of publication decisions or data repositories;

4. peer review of the performance of researchers or research groups in the context of national or international research assessment exercises and awarding scientific or scholarly prizes;
5. peer review in the context of foresight exercises and the development of national or international research agendas.

A key feature of these various practices of peer review is that it is a form of internal control by the scientific community: ‘This process represents the ultimate power exercised by experts who police themselves and who evaluate each other’ (Geisler, 2000, p. 219). This criterion is usually seen as the defining characteristic, based on the widely held assumption that scientists are best equipped to understand the complexities of the research process and results. In its turn, this assumption is the result of the historical process in which modern science developed itself as a fundamentally social process of intellectual, technological and social innovation since the scientific revolution in 17th century Europe (Collins, 1998; Shapin, 1998). Peer review can be characterised as a core family of mechanisms by which the scientific communities control themselves and maintain their social order, scientific ethos and norms (Godlee & Jefferson, 1999; Zuckerman & Robert Merton, 1971).

An important characteristic of peer review is its confidentiality and its distributed nature. Most reviews are conducted by scientists as part of their daily work as knowledge producer. As a result, the more intimate and perhaps most important aspects of peer review are not well known. Also, the art of reviewing is often not systematically included in the training of new research generations.

‘Given the widespread use of peer review, it is surprising that so little is known of its aims or effects although the same might be said of several other, well-established processes of scientific appraisal.’ (Jefferson et al., 2002, p. 2789)

Ethnographic studies of peer review processes have emerged only recently and they tend to focus on peer review in the context of funding decisions (Lamont, 2009). Even as recently as 2013, the lack of robust evidence that peer review is actually the best method the scientific community has as its disposal has been lamented.

‘Surprisingly, especially given that peer review is used by scientists and it is such a fundamental part of researchers’ daily life and career, there have been very few studies aiming at obtaining scientific evidence that peer review is a good way (or even the optimal way) to assess the truthfulness, quality, and potential impact of a scientific contribution or project proposal.’ (Ragone et al., 2013)

In general, however, peer review is still considered the main quality control mechanism due to a lack of serious alternatives (Kassirer, 1994; Smith, 2006).

The multi-faceted nature of peer review has stimulated a diversity of theoretical, empirical, and methodological perspectives from which it has been studied. Although it cannot be denied that intellectual curiosity plays an important role for social scientists, humanists, and scholars in science and technology studies to analyse the peer review processes in more detail, an important characteristic of the literature on peer review is its intimate connection with policy-driven agendas and needs. Indeed the first studies were initiated by anxiety about the integrity and effectiveness of the peer review processes at the US National Science Foundation and the National Institutes of Health (Cole & Cole, 1981; Cole, Simon, & others, 1981; Cole et al., 1978). These studies concluded that peer review results were influenced by a number of different factors, not all of them determined by the scientific quality of the work under review. Since then, peer review has been studied as, among others, a system of codified social norms (Merton, 1973), as the operation of scientific elites in the governance of science at the national and international level (Enders, 2009; Musselin, 2013; Whitley, 1984, 2011), as a process of negotiating epistemic differences (Mallard et al., 2009), as part of the scientific information cycle (Borgman & Furner, 2002; Borgman, 2007), as part of the accumulation of social capital and related credibility cycles (Latour & Woolgar, 1986; Latour, 1987), and as part of the constitution and demarcation of professions (Abbott, 1988; Becher, 1989).

Since peer review is pervasive in both the conduct of research and its management, peer review studies have been conducted in relation to virtually all aspects of science and scholarship. This literature review will focus on the relationship between peer review and bibliometrics of research groups. More general treatises on peer review can be found in Bornmann (2011a); Daniel (1993); Frodeman & Briggle (2012); Godlee & Jefferson (1999); Holbrook (2010); and Weller (2001). The emerging field of career studies is increasingly paying attention to the role of peer review in career promotions, fellowships and scientific prizes (Hicks & Katz, 2011; Mallard et al., 2009; Pezzoni et al., 2012; Van Arensbergen, 2014; Wouters et al., 2010). For studies of the process of journal peer review see Abelson (1990); Akst (2010); Daniel (1993); Jefferson et al., (2002); and Weller (2001). Analyses of editorial processes can be found in, amongst others, Bornmann and Daniel (2010a, 2010b, 2010c); Bornmann & Mungra (2011); Bornmann (2011b); Cabanac & Preuss (2013); Cabanac (2012); Sugimoto & Cronin (2013). A separate body of research studies the quality of journal peer review by comparing bibliometric indicators of journals with peer ratings of those journals. Although this is a form of comparing bibliometric assessment with peer review, it focuses on the quality of journal peer review rather than on the assessment of the quality of research and is therefore not included in this review. A recent overview of the literature on this relationship can be found in Serenko & Dohan (2011).

Alternatives to journal peer review are discussed in Suls & Martin (2009). A review of peer review of grants can be found in Wessely (1998). Recent studies of peer review in the context of grants and research funding are: Frodeman & Briggle (2012); Mutz et al., (2012a, 2012b); Olbrecht & Bornmann (2010); Reinhart (2009, 2010); Van den Besselaar (2012); and Van Leeuwen & Moed (2012). For a comparative assessment of peer review at funding agencies in the US, Canada, the European Union, and the Netherlands see http://csid-capr.unt.edu/.

Many flaws in various aspects of the procedures and practices of peer review have been reported. As a result, many studies claim that peer review is a process whose effectiveness is a matter of faith rather than evidence. According to the former editor of the *British Medical Journal*, a journal which did extensive experiments with and studies of its peer review process, peer review is 'impossible to define in operational terms' (Smith, 2006, p. 178). According to Jefferson et al. (2002), a review of studies of journal peer review, the belief that peer review is the best method we have is untested and uncertain.

For the most recent, and thorough, review of bias in journal peer review see Lee et al. (2013). This literature review critiques the dominant approach to bias in peer review and points out that the identification of perceived bias and shortcomings of journal peer review is also related to the theoretical perspective of the analyst. Hence, not all forms of social influence need to be seen as problematic. The most important weaknesses of peer review discussed in the literature are (Abramo & D'Angelo, 2011; Cole & Cole, 1981; Cole et al., 1981; Cole et al., 1978; Godlee & Jefferson, 1999; Langfeldt, 2004, 2006; Lock, 1994; Overbeeke et al. 1999; Smith, 2006; Wennerås & Wold, 1997):

- it is slow, inefficient and expensive, although most costs are hidden;
- human judgment is subjective – which may however also be seen as a strength (Lee et al., 2013);
- it is not transparent, almost by definition;
- it is inconsistent, sometimes characterised as a lack of inter-rater reliability;
- it is a biased process (e.g. gender bias regarding career decisions, bias against negative studies in publication decisions, bias in favour of prestigious institutes, bias in favour of dominant paradigms);
- its bias is strengthened by the Matthew Effect (Merton, 1988; Merton, 1968);
- the process can be abused (e.g. to block competitors, to plagiarise, to insert abusive comments);

- it is not very good at identifying errors in data or even in detecting fraudulent science (Martin, 1992);
- it cannot process the complete scientific output of a nation and will therefore result in distorted rankings (since rankings are sensitive to the selection of submissions to the assessments);
- it cannot provide information about the productivity and efficiency of the scientific system;
- the selection of peer reviewers may create problems because of a variety of reasons (bias, lack of experts in emerging and interdisciplinary areas, lack of experts due to the speed of growth of research areas, etc.).

At the same time, peer review also has its strengths. Its core strengths are: 1) its foundation in specialised knowledge of the subject, methodology and literature relevant for specific decisions, and 2) its social nature (Lee et al., 2013). This is also the main contrast with decisions based on bibliometric indicators or other formal and mechanical methodologies. It may explain why the recognition of these weaknesses has not led to a call to abolish peer review as the central mechanism of quality control. Rather, the identification of weaknesses has stimulated a series of experiments with different forms of peer review, especially in the area of journal publishing and grant reviews. The most important areas of improvement of peer review are (Jayasinghe et al., 2006; Marsh et al., 2008; Pontille & Torny, 2014; Smith, 2006; Welpe, 2014):

- single blind or double blind peer review to remedy the bias in favour of prestigious institutions;
- post-publication review instead of pre-publication review;
- open peer review to counter the risk of abuse of peer review and increase accountability;
- training of reviewers to improve the quality of the reviews;
- developing new types of peer review (e.g. a focus on methodology rather than substantive quality criteria as developed by PloS ONE in journals or other reader systems in grant review).

The recognition of problematic aspects of peer review has also led to a call to replace or supplement peer review by citation indicators and other metrics that may measure aspects of scientific and scholarly quality and impact (Van Raan, Moed, & Van Leeuwen, 2007).

2.2. Correlating bibliometrics with peer review

The comparison of quality and influence assessments based on peer review on the one hand and bibliometric indicators on the other has been central in the field of bibliometrics from the very beginning. The reason is simple: in the early 1960s when the Science Citation Index was created it was completely unclear what a number like the citation rate might mean (Wouters, 1999). After Eugene Garfield had published his proposal for a citation index in the journal *Science* (Garfield, 1955), the first responses were either negative or a baffled silence. The field of bibliometrics did not yet have credentials. The early bibliometricians therefore set out to empirically investigate to what extent the number of citations correlates with peer judgment of either the quality or the influence of a scientific work. This topic has, in other words, been constitutive for the field and has been an active research theme to the present day. Due to the development and institutionalisation of the citation as a sign of science and an indicator of some form of performance (Bornmann & Daniel, 2008; Nicolaisen, 2007; Wouters, 2014), the role of these studies has been partially inverted. They now also investigate the validity of peer review, rather than of bibliometrics.

In the following section of this review, the focus will be on the more recent studies, although particularly influential older studies will also be included. For older studies see also: Elkana et al. (1978); Moed & Glänzel (2004); Moed (2005a); Nederhof (1988); and Van Raan (1988).

2.2.1. Journal peer review

A recent literature review of peer review from the perspective of bibliometrics confirmed that in general a positive correlation is found between peer judgments of the quality or influence of research groups and institutes on the one hand and their citation impact on the other (Bornmann, 2011a). This is especially true in the case of journal peer review. All five studies covered in this review (Bornmann & Daniel, 2008a, 2008b; Bornstein, 2011; Opthof et al., 2000; Wilson, 1978) confirmed that editorial decisions on acceptance or rejection of journal manuscripts indicated ‘a rather high degree of predictive validity’ (Bornmann, 2011a, p. 122). The low number of studies on this topic is caused, according to the reviewer, by the labour intensity of the studies. It requires detailed information about all submitted manuscripts, including the rejected ones, and their subsequent citation counts. In addition, many journals may be reluctant to open their archives for these types of studies (if they have these archives). A positive correlation between journal peer review and citation impact has been confirmed in Benda & Engels (2010); Bornmann & Daniel (2010c); and Cicchetti (1991).

2.2.2. Grant peer review

With respect to grant peer review, the literature is somewhat less clear. The most recent literature review on grant scientific peer review (Bornmann, 2011a) found only six studies on the assessment of citation counts as predictive for decisions regarding grants and fellowships. The studies analysed to what extent applicants whose proposals were funded were cited more frequently than their less successful colleagues. The results are contradictory. Some studies find a positive correlation between funding and citation impact, other studies question whether grant peer review and citation impact are correlated.

Armstrong et al. (1997) reported on a 10 year study of researchers funded by the Heart and Stroke Foundation of Canada. Funded individuals published more papers and they received more citations than their unfunded comparison group. Since causality can operate in both directions, no conclusions regarding the cause of the successes were drawn. Bornmann et al. (2008) analysed two programmes of the European Molecular Biology Organization to support the best post-doctoral fellows and young group leaders in the life sciences. Funded researchers score higher on number and citation impact of their research publications after their application for the funding. At the same time, between 26% and 48% of the decisions made to award or reject an application show either a type I or type II error. In other words, the correlation is not perfect and causality has not been shown. Bornmann & Daniel (2005, 2006) analysed the selection procedure of young researchers implemented by the Boehringer Ingelheim Fonds (BIF), a foundation for the promotion of basic research in biomedicine on reliability (agreement among peers), fairness (lack of bias) and predictive validity (selection of the best researchers). Reliability was high: the reviewers agreed in 76% of the cases. The study found no bias related to gender, nationality, field of study or institutional affiliation for post-docs. In contrast, it found evidence of bias related to gender, field of study or institutional affiliation in doctoral fellowship decisions (no nationality bias was found). Funded researchers were cited more often than the average paper in the set of journals of the relevant fields (both before and after they were funded). The authors conclude that the selection procedure is successful in selecting the best researchers.

Hornbostel et al. (2009) studied a comparable funding programme aimed at young researchers at the German Research Foundation (DFG). The reviewers tended to select applicants with slightly higher citation rates and, in particular, successfully identified highly productive young researchers. However, participation in the programme had not decisively influenced research performance in the examined fields of medicine and physics. In medicine, in particular, no differences between approved and non-approved applicants could be found a few years after the funding decision.

Cabezas-Clavijo et al. (2013) analysed the relationship between peer-based funding decisions and bibliometric indicators for Spanish researchers in 23 research areas in the context of the Spanish national R&D Plan 2007. The differences between authors of rejected and accepted proposals were measured bibliometrically. Bibliometric indicators for applicants of accepted proposals showed a better previous performance than for applicants of rejected proposals. The number of published articles and the number of papers published in journals that belong to the first quartile ranking of the Journal Citations Report are the indicators that best explain the grant decisions. However, the correlation between peer review and bibliometric indicators is heterogeneous among most areas.

Social sciences and education are the only areas where no relation was found. The study suggests that apparently funding decisions are usually taken on the basis of the past performance of the principal investigator. In addition, peers may tend to rate researchers positively on the basis of their background rather than only on the basis of the submitted proposals. In other words, peer review and bibliometrics cannot be treated as fully independent of each other, since reviewers may use bibliometric data in their evaluation.

Van Leeuwen and Moed (2012) analysed the correlation between funding and citation impact in the fields of mathematics, astronomy, chemistry and geosciences. The comparison was made with the applicant's 10-year publication oeuvre (the funding source of this research was not taken into account). The study found that the three funding councils tend to attract research proposals from the better groups in the fields they cover. The applicants whose submitted proposals were granted – and the research groups they represent – tend to generate a higher citation impact at their international research fronts than those whose submissions were rejected.

A different result was obtained by Van den Besselaar and Leydesdorff (2009). This study analysed the funding by the Dutch science foundation of social and behavioural research (NWO MaGW). The conclusion of that study is that the science foundation funded researchers with a good/excellent performance, particularly when compared with all the unfunded researchers. However, the council was not able to distinguish between the good/excellent researchers that got funded and the next group of unfunded researchers of roughly the same volume. This leads to the conclusion that NWO MaGW is not able to select the ‘best’ researchers. The methodological design of this study was questioned by peers, as a result of which the implications of the study are not yet completely clear (Van Leeuwen & Moed, 2012).

It must be emphasised that the study of grant peer review meets the same challenges as the study of journal peer review. Often the data are not available or not accessible, and the required data collection design is very labour intensive. According to Bornmann (2011a), there are moreover specific methodological problems due to which the different studies are not always comparable. Studies of the citation impact in relation to funding run the danger of circular reasoning. This is especially the case when successful and unsuccessful applicants are compared. Higher citation impact for the successful researchers after they obtained funding may, after all, be the consequence of the funding. It can therefore not be concluded that they were the better researchers to start with. To circumvent this type of circularity, Bornmann (2011a) advocates the use of discipline-specific citation reference standards that are independent of the funding decisions (e.g. citation distributions or averages for the field as a whole). The same holds for studies of fellowship and career progression in relation to bibliometric performance indicators.

2.2.3. Correlating bibliometrics and peer judgment of research groups

Nederhof and Van Raan (1993) analysed the relationship between bibliometric indicators and peer review in an interactive experiment funded by the UK Economic and Social Research Council (ESRC) in which the peer judgment and the citation analysis were performed simultaneously. The scientific performance of six research groups in economics were measured and two experts in economics were asked to rate these groups both before and after the scientometric analysis had been done. The peers were also asked to comment on the citation analysis. Peer judgments and bibliometric findings were generally in agreement. The peers found the bibliometric analysis a useful check on peer review. Nevertheless, the authors conclude that 'excessive reliance' on these measures needs to be avoided:

'Research groups ought not to be encouraged to think that their rankings will be closely dependent on citation rankings, since the resulting re-direction of research and publication efforts might well be at the expense of the goals for which a research programme was established in the first place.' (Nederhof & Van Raan, 1993, p. 366)

Rinia et al. (1998) analysed 56 research programmes in condensed matter physics in the Netherlands. This set of research programmes led to more than 5,000 publications and nearly 50,000 citations. The study showed varying correlations between different bibliometric indicators and the outcomes of a peer evaluation procedure. At the level of teams the strongest correlation between peer judgment and bibliometric indicators was found. Correlations proved to be higher for groups which were involved in basic science than for groups which were more application oriented.

Positive associations were also obtained by Meho and Sonnenwald (2000). This study analysed the relationship between citation ranking and peer evaluation in assessing senior faculty research performance. It applied citation context analysis as well as book review content analysis in addition to the evaluation data based on peer review. The normalised citation ranking and citation content analysis data yielded identical ranking results. Normalised citation ranking and citation content analysis, book reviews, and peer ranking were highly correlated for high-ranked and low-ranked senior scholars. The study concluded that additional evaluation methods and measures that take into account the context and content of research appear to be needed to effectively evaluate senior scholars whose performance ranks in the middle.

Lewison (2001) analysed the relationship between citation analysis and peer ratings of books in medical history and found a high degree of agreement about what the best books were.

Aksnes and Taxt (2004) studied the relationship between bibliometric indicators and peer review outcomes at a Norwegian university. They found positive but weak correlations. The study attributes this weak correlation to shortcomings in the peers' assessments, and in the indicators, as well as in a lack of comparability. This points to a general conclusion that can be drawn from the literature: the imperfect correlations between bibliometric indicators and peer review can partly be explained by variation in qualitative peer-based judgments.

Van Raan (2013) presented an analysis of the statistical correlation between the h-index and several standard bibliometric indicators, as well as with the results of peer review judgment. The study was based on an evaluation study of 147 university chemistry research groups in the Netherlands covering the work of about 700 senior researchers during the period 1991–2000. The results showed that the h-index and the normalized citation impact indicator from the Centre for Science and Technology Studies (see section 1.2.3.1) both relate in a quite comparable way with peer judgments. However, for smaller groups in fields with ‘less heavy citation traffic’ the h-index seems to be a less appropriate measure of research performance.

2.2.4. Bibliometrics and national research evaluation exercises

The rise of national assessment exercises that are based on peer review in the UK, Italy, the Scandinavian countries, the Netherlands, Belgium, and a number of other countries has created a novel opportunity to studying to what extent the outcome of peer review evaluations can be predicted by, or is correlated to, assessments that are only based on bibliometric data. It has resulted in a body of knowledge about the ways in which bibliometric rankings and assessments relate to peer review-based rankings and assessments. The choice of bibliometric indicators varies by study. In general, the study designs differ considerably in many respects.

Derrick et al. (2011) analysed the relationship between peer judgment of a researcher's influence on the one hand and a range of citation metrics on the other in six fields of public health in Australia. Four of the six fields displayed a moderate positive correlation. Two fields showed no relationships or negative relationships (tobacco and injury research). The authors conclude that in the latter cases, researchers are evaluated by their peers on other criteria than visibility in the literature. The study therefore advises the combined use of metrics and peer review. In addition, they conclude that the most appropriate metrics for research evaluation will differ between research fields. As we have seen, this is a recurring theme in the literature.

In a series of articles Abramo and his team compared the university performance ranking lists from the first peer-review Italian research assessment exercise (VTR 2006) (Franceschet & Costantini, 2011) with those obtained from evaluation simulations conducted with bibliometric indicators (Abramo, Cicero, & D’Angelo, 2011, 2012a, 2012b; Abramo, D’Angelo, & Cicero, 2012; Abramo, D’Angelo, & Di Costa, 2011a, 2011b; Abramo, D’Angelo, & Viel, 2011; Abramo & D’Angelo, 2011). The studies present perhaps the strongest plea available in the literature on peer review and bibliometrics to replace peer review-based assessment by metrics-based assessment for the hard sciences. It should be noted that the studies are based on the assumption that citations are a good representation of quality in these fields (Abramo, Cicero, & D'Angelo, 2012a). According to the study the bibliometric measurements are better able to assess quality in the 'natural and formal sciences' than the VTR 2006 exercise.

An important feature of this set of studies is the creation of a national database of all publications by Italian researchers on the basis of WoS and Scopus data. The database lists all scientific publications produced since 2001 (about 272,000 articles and reviews, and 100,000 conference proceedings) by all public research organisations in Italy (approximately 350 in total). The database attributes publications to every academic author with an error of less than 5% (Abramo & D'Angelo, 2011). This dataset includes approximately 95% of the products presented to the VTR for the natural and formal sciences, and a few fields of social and economic sciences, i.e. for the output of 70% of the total research staff of Italian universities. The authors conclude that the database is 'highly representative' of the entire national research output for the natural and formal sciences. Based on the database, an evaluation support system has also been developed, with the potential for producing rankings according to a number of performance indicators (productivity, productivity weighted for quality, productivity weighted by the number of co-authors, by the order of the author's co-listing, etc.). These indicators can be applied to measure the performance of each Italian university researcher active in identified fields of research. The decision support system based on the database does not require any input by the research institutions under observation. According to the authors, this offers savings in indirect costs and time for execution of evaluations. The lower costs would permit greater frequency of evaluations: on the order of months, rather than years. The authors acknowledge that the error rate in the data could result in distortion at the level of the individual researcher. In their view, the database is not an evaluation system itself, but a support system analogous to diagnostic imaging tools in medicine. For evaluation of small units, further checking is always desirable. The authors propose that similar databases should be developed in other countries, thereby enabling international comparison.

The studies by Abramo and his colleagues compared peer review and bibliometrics in terms of: accuracy, robustness, validity, functionality, time and costs (see section 2.1, p. 41). The main argument in this set of studies in favour of bibliometrics over peer review (including metrics-informed peer review) is the capability to measure all output. For the national assessments, institutes can avoid the submission of selected publications, resulting in less error and cost savings. The Italian peer-based VTR 2006 assessment processed 9% of the total output of the country. According to Abramo et al. this shows that the peer-based national assessments cannot be robust, nor can they be valid. Because bibliometrics can measure the complete national scientific output, measures of productivity can be developed, which is not possible in any peer review-based assessment. In addition, the metrics-based rankings will not be distorted by the selection of submissions (a disadvantage of peer-based rankings which can by definition only assess a small part of the total scientific production). The weakness of bibliometrics identified by Abramo et al. is that it cannot be applied to all scientific outputs but only to journal publications, conference proceedings and patents.

To sum up, this set of studies on the Italian research system concludes that the superiority of bibliometrics over peer review is 'evident for the natural and formal sciences', along the dimensions of:

- robustness: bibliometrics allows evaluation of all, rather than a subset of overall output;
- validity: it avoids any distortions that could occur during internal selection of products to be evaluated;
- functionality: in providing evaluations for single scientists, then proceeding step by step to research groups, and ever larger aggregations, it permits each institution to allocate resources in an efficient manner;
- cost and time effectiveness: it provides a dramatic saving on direct and indirect costs, and dramatically reduces time of execution (Abramo & D'Angelo, 2011, p. 510).

Gambardella et al. (2013) analysed Italian research in economics, business and statistics (12,000 publications dated 2004-2010). A random sample from the available population of journal articles shows that informed peer review and bibliometric analysis produce similar evaluations of the same set of papers. The study concludes that these two approaches are substitutes, either because of independent convergence in assessment or because of the influence of bibliometric information on the community of reviewers.

2.2.5. Predicting the outcomes of the UK RAE and REF by bibliometrics

The creation of a national database of the RAE 1992 was used to analyse the correlation between scientometric indicators and the RAE outcomes in the field of business and management studies (Taylor, 1994). The indicators used were: department size, number of articles in refereed journals, number of research students, and research income. The study found 80% correlation in a regression model based on these indicators. The authors conclude that these indicators provide useful additional information but cannot be used to replace peer evaluation.

Oppenheim (1995) performed a citation analysis of all 217 academics who teach in UK library and information science schools. The results were ranked and the ranking was compared with the results of the RAE 1992. The study showed a statistically significant correlation between the numbers of citations received by a department in total, or the average number of citations received in the department per academic, and the RAE rating. The paper concludes that this provides support for the validity of citation counting, even when using just the first authors as a search tool for cited references. The paper also concludes that the cost and effort of the RAE may not be justified.

A positive correlation between peer and bibliometric rankings was also found in a later study of the field of business and management studies (Thomas & Watkins, 1998). Institutions were rated on the basis of the bibliometric performance of the journals in which they published. Although they propose to use their indicator in the context of the RAE, the authors do not conclude that this bibliometric indicator could replace peer review in its entirety. They still see a role for peer review for emergent work that is not yet visible in bibliometric indicators and for high quality work that is published in less prestigious journals. The main advantage of the indicator is seen in its neutrality with respect to the submission strategy of the higher education institution.

A large-scale study of the RAE 2001 analysed all 203,743 individual submissions (Mahdi et al., 2008). Citations were counted at the level of the individual submissions for all journal articles covered by WoS (55% of the submissions). The citation counts were a reasonable proxy for the RAE results in the biological sciences, clinical sciences, chemistry and psychology. They correlate much more weakly with the RAE results for a large number of disciplines, including fields in the biomedical and engineering sciences, and including fields that are well covered in WoS. The citation counts are even less valuable for fields not well covered in the citation index, according to the study.

Norris and Oppenheim (2010a) examined whether the h-index and its variant the g-index (see section 1.2.1) correlate with the outcome of the UK RAE 2008 rankings. The study measured the collective h-index of the submitting departments. Three units of assessments were measured: library and information science, anthropology, and pharmacy. The results were mixed. In the field of pharmacy, a strong correlation existed between the RAE ranking and the median bibliometric scores. Library and information science showed a moderate correlation, whereas in anthropology the correlation was negative or non-existent. The fact that anthropology is less well covered by WoS does not explain this result. In earlier studies, the same authors had analysed the correlation between bibliometrics and RAE results in the field of archaeology, which is even less well covered by WoS than anthropology. In these studies, they found a strong correlation between the RAE and bibliometric indicators (Norris & Oppenheim, 2003; Oppenheim, 1997). The different outcomes are probably best explained by the different designs of the studies. The earlier work measured total and average citation counts (rather than the h-index). In addition, Norris and Oppenheim (2003) had access to the data about submitting researchers at the individual level. The different designs of the research assessment exercises may also explain part of the differences.

Clerides et al. (2011) analysed the same correlation for the field of economics. The study posed the question of how different the outcome of the RAE would be if it had relied exclusively on performance indicators. The analysis showed that the RAE rankings cannot fully be explained by the bibliometric indicators. The study concludes that this shows the ‘discretion exercised by the panels’.

Motivated by the question as to whether bibliometrics can provide a low cost alternative to the seemingly expensive RAE peer rankings, Butler and McAllister (2011) applied a metrics-based model developed in earlier studies of political science to the field of chemistry, using data from the RAE 2001. The model identified the best predictors of the RAE results in political science (Butler & McAllister, 2009). The most important and statistically significant predictive variables were: citations (including citations to journal articles, books and book chapters) and departmental size (represented by student numbers). Research income was not a strong predictor, nor was departmental size if measured by staff numbers. The study also found that whether or not a department had a member on the assessment panel was a strong predictor of the RAE outcome. This model was able to explain 60% of the variance in RAE outcome in political science, and 86% of the variance in chemistry. However, in political science citation count was the best predictor, while in chemistry research income correlated strongest with the RAE results. The latter indicator had almost no predictive power in political science. The study also analysed in what sense the RAE results would have been different if the model had been applied. For 34% of the 113 departments, the results would have been different if the RAE had been purely metrics based. Taken as a whole, the rankings were very close to each other. The study concludes that a metrics-based model, using indicators drawn from a range of readily available measures, will yield results close to those of a peer-based evaluation model, and can be used with confidence. However, the results also point to strong differences between SSH disciplines and STEM disciplines. No single model can be used across all disciplines. Any metrics approach to performance evaluation has to use a discipline-specific suite of indicators. A STEM ‘basket’ might rely almost exclusively on citation and external income data, with little need for other indicators. For these disciplines, the focus would be on selecting citation measures that are sensitive to varying citation practices within broad disciplines, that do not militate against the inclusion of recent publications that have had little time to attract citations, and that do not prove to be a disincentive for collaborative and interdisciplinary research activities, according to the study. In the humanities and social sciences, such a basket may be more problematic. The cost of tailor-made indicators may rise and peer review may in the end turn out to be necessary, even in ‘a stream-lined metrics-based approach’, the authors speculate.

Taylor (2011) studied the extent to which the outcomes of the RAE 2008 can be explained by a set of quantitative indicators in business and management, economics and econometrics, and accounting and finance. The main finding was that each of the three components of research activity (namely, research output, esteem and research environment) was highly correlated with quantitative indicators.

The judgment of the panels was biased in favour of Russell Group universities. The study also showed some evidence of bias in the judgments of the economics and econometrics panel. The authors support the use of quantitative indicators in the research assessment process. They propose in particular a journal quality index. ‘Requiring the panels to take bibliometric indicators into account should help not only to reduce the workload of panels but also to mitigate the problem of implicit bias.’ This study is therefore an example of the approach to counter bias in peer review with the help of bibliometrics.

In the field of social work and social policy and administration, a metrics-based assessment can predict reasonably well the overall outcome of the RAE 2008 in terms of research environment, but not in terms of research outputs (McKay, 2012). The study showed that authors did not always choose to submit their most highly cited work to the RAE. It also showed that it is possible to explain a great deal of the variation in scores awarded for research environment, ‘but rather more difficult to find a quantitative counterpart to the peer assessment of research outputs’. According to the study, this supports the panel’s insistence on reading the particular works, rather than using shortcuts based on the identity of the journal. The output measures applied in the study related to the type of output and their journals and publishers. ‘At least in this subject, metrics are more suited as handmaiden to peer review than its replacement’, the study concludes.

Kelly and Burrows (2012) developed an exploratory model to predict the RAE 2008 in the field of sociology. They found that 83% of the variance in outcomes can be predicted by a combination of simple metrics: the quality of journals in the submission, research income per capita and the scale of research activity. The most powerful single predictor of how well a submission did in the RAE 2008 was how well it did in the RAE 2001. The model used a sophisticated indicator to measure both the normalised citation rate and the centrality of a journal for the field. Measured in this way, the percentage of journal articles included in a submission that were published in the ‘top’ quartile of journals proved to be the best citation-based predictor of the outcomes of RAE 2008.

It should be noted that the study restricted itself to those submissions that involved publications in journals covered in Thomson Reuter's Journal Citation Reports (34% of all submissions). The study discusses the irony involved:

‘Here we have a series of decisions made by sociologists who inhabit an intellectual world generally dominated (in the UK at least) by subjectivism, anti-positivism and relativism that are themselves largely predicted by a set of quantitative indicators formed into a simple regression model embodying what some might view as an opposing ontology and epistemology.’ (Kelly & Burrows, 2012, p. 147)

They conclude that measuring the value of sociology involves ‘multiple mutual constructions of reality within which ever more nuanced data assemblages are increasingly implicated’. They suggest that the model is able to mimic the judgments of the panel as well as it does because the variables are originally derived from qualitative peer review translated into quantitative metrics.

In a study of the relation between RAE 2008 results and reputations rankings, Allen and Heath (2013) found that the number of articles in top journals (defined by the rankings) as a percentage of all submissions was associated with the institution's top grading in the RAE (4*). The types of output also mattered. Top press monographs were most strongly associated with 4* grades. The proportion of articles in top-10 journals also had a positive and significant association with 4* work. The study concludes that publisher reputations are good predictors of research quality as graded by the RAE. The explanation is that the panels are based in the same communities that review manuscripts. Moreover, in many fields publication decisions are influenced, if not steered, by the goal to publish in high-impact journals. If the correlation did not exist, it would mean that the panels were not representative of the discipline as a whole, the authors argue. The study also found that large departments that submitted many outputs did better in the RAE on average. Having a member on the RAE sub-panel also contributed to the score. The study does not make clear whether this can be attributed to bias in the RAE or to the fact that panels are drawn from the departments with high research performance. Overall, the study concludes that both RAE judgments and reputational rankings are based on peer review. The advantage of using the latter is that they are based on the opinion of more people.

Eyre-Walker and Stoletzki (2013) compared three methods of assessing the merit of a scientific paper: subjective post-publication peer review, the number of citations gained by a paper, and the impact factor of the journal in which the article was published. There are moderate statistically significant correlations between assessor scores, when two assessors have rated the same paper, and between assessor score and the number of citations a paper accrues. However, assessor score depends strongly on the journal in which the paper is published, and assessors tend to over-rate papers published in journals with high impact factors. After control for this bias, the correlation between assessor scores and between assessor score and the number of citations is weak. The authors suggest that scientists have little ability to judge either the intrinsic merit of a paper or its likely impact. The study confirms moreover that the number of citations that a paper receives is a stochastic process. The journal impact factor is likely to be a poor measure of merit, since it depends on subjective assessment. The authors conclude that the three measures of scientific merit considered here are poor: ‘in particular subjective assessments are an error-prone, biased, and expensive method by which to assess merit’. The study concludes that notwithstanding its poor performance, the journal impact factor may be the least unsatisfactory measure of merit.

This conclusion is contested by Eisen et al. (2013), although they agree that the study shows that the current practices of research assessment are neither reliable nor consistent. They argue that any single metric that is highly variable is going to pose a problem for research assessment. This is compounded when assessments are based on subjective opinion or other very biased measures, such as the journal impact factor. They propose using not any single indicator, but a suite of metrics at the level of the article. In such a system, they argue, it will also be important to enable research into new metrics of assessment. Crucial to this is the availability of data about research assessment itself.

Mryglod et al. (2013) analysed the correlation between peer review scores in a range of academic disciplines from natural sciences to SSH. The analysis was conducted at the level of the research group. The citation measure in Thomson Reuters Research Analytics was poorly related to RAE scores. The study concludes that these indicators should not be used in place of peer review. However, a measure of total impact in which the size of the department was taken into account, strongly correlated to overall strength according to the RAE in a number of fields. This is especially the case for large research groups. For smaller groups the correlation becomes weaker or disappears (which is due to the smaller size of the data sets). This correlation is moreover stronger for the hard sciences than for other fields. In the more specific comparison of academic impact and quality, the study finds weak correlations for the majority of disciplines: chemistry, physics, engineering, geography and environmental studies, sociology and history. The authors conclude that citation indicators should not be used in isolation to compare or rank research groups or higher education institutes.

Mryglod et al. (2014) aim to predict the results of the REF 2014 outcome on the basis of the h-index of departments. In a follow-up study, the prediction is compared with the actual results of the most recent REF exercise (Mryglod et al., 2015). The study first analysed whether two institutional bibliometric indicators correlated with the results of the RAE 2008. The conclusion was that a version of the departmental h-index was a better predictor of the peer review ratings than the institutional normalised citation impact. On this basis, Mryglod et al. (2014) predicted the outcome of the REF 2014 peer review exercise by calculating the departmental h-indices for the publication period 2008-2014. The publication of the results of the REF 2014 enabled comparison of this prediction with the actual results, and these results are negative (Mryglod et al., 2015). The correlations between the departmental h-index and the REF 2014 outcomes were approximately as strong as the correlation between that h-index and the earlier RAE 2008 results. However, the similarities are still not good enough:

‘...h-indices used in this way do not track the peer review exercises well enough for them to form a component of, or substitute for those exercises. Additionally, we found very poor correlations between the predicted and actual changes in the ratings. This means that the departmental h-index does not offer a way to foretell the direction of changes of universities in the rankings in these subject areas.' (Mryglod et al., 2015, p. 4).

The authors draw implications for the use of metrics in research assessment exercises as follows:

Our study shows that a very different landscape would have emerged in the UK if REF 2014 had been entirely and simplistically based on the automated departmental h-index. A wise academic subject expert can, perhaps, use such a metric to gain perspective in combination with other approaches, taking into account many nuances such as scientific context, subject history and history of science generally, technical aspects, future perspectives, interdisciplinarity, and so on. Clearly, however, over-reliance on a single metric by persons who are not subject experts could be misleading, especially in increasingly managed landscapes in which academic traditions are diminished or eroded.' (Mryglod et al., 2015, p. 5)

2.3. Informed peer review

Although no complete consensus exists in the bibliometric literature about what citation indicators and bibliometric measures exactly mean, the vast majority of bibliometric experts see citations as a proxy measure for impact of the work on the relevant scientific communities (Martin & Irvine, 1983; Narin, 1976; for an overview of citation theory see Nicolaisen, 2007). The act of counting citations abstracts from the substantive information in the scientific literature and is based on the formal relationships among references and citations. As a consequence, by definition many forms of peer review cannot be replaced by bibliometric indicators. This is clearly the case where the research has not yet been published, but also in many instances where substantive judgment is required (for the limitations of bibliometric indicators see section 1.2).

This has given rise to the concept of informed peer review (Butler, 2007; Moed, 2007). The basic concept is that a judicious application of specific bibliometric data and indicators may inform the process of peer review, depending on the exact goal and context of the assessment (Nederhof & Van Raan, 1993; Nederhof, 1988; Van Raan, 1996). Informed peer review is in principle relevant for all types of peer review and potentially at all levels of aggregation, although it is often seen as especially relevant at the micro and meso level of peer review. To what extent large-scale research assessment exercises should be based on peer review (whether informed by metrics or not) rather than on metrics is a contested issue in the literature.

According to Moed (2007), based on an analysis of the UK RAE exercise, the future of research evaluation will be based on forms of informed peer review. The central thesis of the paper is that the future of research evaluation rests with an intelligent combination of advanced metrics and transparent peer review. It argues that metrics, especially a sophisticated type of citation analysis, may provide tools to keep the peer-review process honest and transparent. Both metrics and peer review have their strengths and limits. The challenge is to combine the two methodologies in such a way that the strength of the first compensates for the limitations of the second, and vice versa.

‘Outcomes of citation analysis must be valued in terms of a qualitative, evaluative framework that takes into account the substantive content of the works under evaluation. This can be done by peers only. The conditions for proper use of bibliometric indicators at the level of individual scholars, research groups or departments tend to be more readily satisfied in a peer-review context than in a policy context. It can therefore be argued that bibliometric analyses at such lower aggregation levels normally best find their way to the policy arena through peer assessments.’ (Moed, 2007, p. 577)

Butler (2007) developed a model for a ‘balanced approach’ to research evaluation on the basis of the experiences in the UK and Australian assessment exercises. According to Butler (2007), too often new ways of research assessments are being proposed in a cyclical way without learning from past experiences. The study points to the tendency in research policy to see the measurable part of scientific quality as a proxy measure for quality in total. However, it notes that most informed researchers see bibliometric information as valuable not to replace peer judgment, but to make the latter debatable, to provide additional information, and to help peer review to become more transparent. Indicators can also be useful in cases of doubt and they can be used to highlight gaps in the knowledge of peers. The study also mentions the danger of haphazardly combining indicators with qualitative information. The indicators need to be selected carefully in order for the combination to be meaningful. This means that a ‘suitable suite’ of indicators needs to be defined, and they will tend to vary by field.

As concrete example of forms of informed peer review in practice, the Dutch Standard Evaluation Protocol has been mentioned (Moed, 2005a, p. 233). Although the evaluation committees are free to use or not use citation data, in many disciplines it has become common practice to have the evaluation informed by a citation analysis report (Colwell et al., 2012; KNAW, 2010, 2011; Phillips, 2012). Due to the confidential nature of peer review, it is actually not clear to what extent citation data are informing, and thereby influencing peer review processes. However, since bibliometric data have become widely available, it is highly likely that some form of use of these data and indicators has become common sense in many fields and contexts (see section 1.3).

Whereas Moed (2007) and Butler (2007) argued in favour of informed peer review on theoretical grounds and on the basis of the studies showing positive correlations between outcomes of peer judgment and citation data, other studies have developed additional proposals regarding, and arguments in favour of, informed peer review.

An empirical argument in favour of informed peer review is developed in Donovan (2007). The study argues that quantitative indicators are as infused with human values as are qualitative approaches. It shows that quality and impact metrics have followed a trajectory ‘away from the unreflexive use of standardised quantitative metrics divorced from expert peer interpretation, towards triangulation of quantitative data, contextual analysis and placing a renewed and greater value on peer judgment combined with stakeholder perspectives’. In other words, the study seems to suggest that a trajectory towards the fuller development of informed peer review is emerging.

A specific argument in favour of informed peer review is the limitation of citation analysis to predict future work. Mazloumian (2012) tested the assumption that citation counts are reliable predictors of future success, analysing complete citation data on the careers of approximately 50,000 scientists. Their results show that among all citation indicators, the annual number of citations at the time of prediction is the best predictor of future citations. In addition, future citations of a scientist's published papers can be predicted accurately. However, future citations of *future* work are hardly predictable. On this basis Penner et al. (2013) warn that the impact of papers published in the past does not necessarily correlate with that of papers published in the future. This can be seen as an argument in favour of (informed) peer review to assess the potential of future work by researchers.

Informed peer review may also be used to provide feedback on the design of new performance indicators. In fact, this is the default form of validation of scientometric analyses by the relevant scientific communities under measurement (documentation of these interactions is often provided in the acknowledgements or methodological sections, not by citations). In a project funded by the Rectors’ Conference of the Swiss Universities to develop performance indicators for the humanities (Hug et al., n.d., 2013; Ochsner et al., 2012, 2014), the links between quality criteria used or recognised by humanities scholars and possible indicators was explored in the fields of German literature studies, English literature studies, and art history. The study concludes that humanities researchers’ refusal of evaluations could be alleviated if the assessment is based on peer review using consensual criteria and aspects for research quality and if the scholars are involved in the process early on (i.e. definition of research quality). Indicators that are linked to the humanities scholars’ notions of quality can be used to support peers in the evaluation process.

Bibliometrics has also been used to propose improvements in the system of large-scale peer review. A recent study of 10 different conferences in computer science (ca. 9,000 reviews on ca. 2,800 submitted contributions) has explored possible improvement of conference peer review (Ragone et al., 2013). The study aimed to establish scientific evidence that peer review works (or that it does not work). The data analysis did not provide a definite answer to this research question. It did however formulate a number of ways to improve the peer review process. The study found ‘a significant degree of randomness in the analysed review process, more marked than we expected’. The study finds a low correlation between peer review outcome and later citation impact of the accepted contributions. In addition, the assessment scale influenced the marks given by reviewers: the study provides evidence of systemic rating bias, with reviewers constantly giving lower or higher marks than all other reviewers. The study developed statistical approaches to influence some parameters of the peer review process (such as the number of papers given to each reviewer) in order to improve the overall quality of the peer review system. According to the authors, this statistical model can be used as a decision support system in state-of-the-art editorial management systems (this includes robustness analysis, disagreement analysis, band agreement analysis, bias analysis, un-biasing procedures, a-posteriori validity analysis with respect to specific target parameter(s), a-posteriori or on-the-fly marks accuracy evaluation, as well as statistical approaches to tune review process parameters). The authors suggest that this may better inform the conference and panel chairs. This way of using metrics to support, rather than replace, human judgment has not been explored very often in the literature.

Informed peer review can also be used to increase the degree of participation in review processes by non-academic stakeholders. Derrick and Pavone (2013) reviewed government policies from three countries (the UK, Australia and Spain). Each country is committed to the democratisation of science for policy while this commitment does not play a role in research evaluation policies. The study argues that this discrepancy must be addressed, which shows the ‘inherent utility of bibliometrics for translating bottom line information to non-academic stakeholders’. When used in combination with appropriate peer-review methods, bibliometrics, as part of an informed peer-review process, has the potential to widen scientific participation by allowing non-academic stakeholders to access scientific decision-making. ‘While evaluation by peers ensures that scientific work is evaluated competently, it confines scientific evaluation to a minute, hyper-specialised committee of “experts”, operating behind closed doors’, the study argues. It provides a table with specific advantages and disadvantages of peer review versus bibliometrics, which shows that peer review and bibliometric may complement each other in virtually all aspects. An advantage of metrics is that manipulation can more easily be identified. It can also analyse the complete output of a nation (Abramo & D’Angelo, 2011). In addition, their formal integration is preferable over the current model where bibliometrics, or at least the perceived reputation of a journal, are applied ‘haphazardly without guideline or benchmarks to ensure the transparency, validity and accountability of evaluation outcomes’ (Derrick & Pavone, 2013). The study proposes a change in dialogue from ‘whether bibliometrics should be used to how they should be used in future evaluations’.

This topic is part of an emerging literature in which the ‘relevance gap’ (Nightingale & Scott, 2007) or the ‘evaluation gap’ (Wouters et al., 2010) are seen as a major shortcoming of the traditional disciplinary peer review systems and practices. Although this literature is mainly in the area of science policy and higher education studies and not itself mainly concerned with bibliometrics, it may provide an important theoretical context for further developing concepts of informed peer review and informed expert review (Etzkowitz & Leydesdorff, 2000; Hemlin & Rasmussen, 2006; Martin, 2011; Nowotny et al., 2001). The decades of experience among bibliometricians with bibliometric databases, indicators, and deliberations with users of these data and indicators (sometimes under the flag of 'validation of bibliometrics') may prove to be more useful in a wider sense than has previously been realised.

This development has recently become more pronounced by the increased need for guidance in the use of bibliometric and other performance indicators. A recent analysis of the bibliometric literature has shown an increased role of authors not affiliated within the bibliometric discipline (Jonkers & Derrick, 2012). In response to the growing availability of bibliometrics and to concerns about the potential for abuse and unintended effects, the bibliometric community started a number of initiatives to initiate principles of good evaluation practices, building on initiatives in the community that started decades ago (Moed & Glänzel, 2004; Noijons & Wouters, 2014; Wouters et al., 2013). Perhaps a new body of 'translational bibliometrics' literature to flesh out the concept of informed peer review will emerge from these initiatives.

2.4. Peer review and bibliometrics conclusions and summary

Peer review is a general umbrella term for a host of expert-based review practices that show considerable variation. The most important forms of formalised peer review are journal review of manuscripts, peer review of applications for funding and career promotions, and national peer review-based research assessments.

The results of peer review-based decisions generally show positive correlations to selected bibliometric performance data. However, it matters a lot exactly which forms of peer review and which specific dimensions of peer review are being related to exactly which bibliometric indicators. It is also important to define exactly how these bibliometric indicators are being measured and on the basis of which data sets. Bibliometric measures ought not by definition to be seen as the objective benchmark against which peer review is to be measured.

The literature on the relation between peer review and bibliometrics has not (yet?) developed a common methodology. As a result, different outcomes of studies with respect to the strength of the correlation between peer review and bibliometric measurement may be caused by different research designs. Many studies of the relation between bibliometric performance and funding decisions that report a positive correlation are plagued by circular reasoning: the better citation performance of funded researchers may very well be the result of this funding.

The popular view that citation rate is a measure of scientific quality is not supported by the bibliometric expert community. Bibliometricians generally see the citation rate as a proxy measure of scientific impact or of impact on the relevant scientific communities. This is one of the dimensions of scientific or scholarly quality. Quality is seen as a multidimensional concept that cannot be captured by any one indicator. Moreover, which dimension of quality should be prioritised in research assessments may vary by field and by mission.

The literature shows varying strengths of correlation between bibliometric indicators and peer review assessment. Correlation strengths vary between fields both within the natural sciences, the social sciences, and the humanities. It may even vary within fields. In some fields citation-based indicators are strong predictors of peer review outcomes, in other fields this may be research income, and in a number of fields there is no correlation. In general, the correlation between bibliometrics and peer review is weaker in most fields in the humanities, the applied fields, the technical sciences, and the social sciences. This is partly caused by less coverage in the citation databases, but also by varying citation and publication cultures.

Peer review and bibliometric data are not completely independent. Citation data are in the end based on scientists who cite or do not cite particular publications. The same communities are the source of the peer review data. Although the meaning of the citation cannot be deduced from the role of the literature reference, it does explain the strongly to moderately positive correlation between peer review and bibliometrics. In addition, peer review decisions may have been influenced by prior knowledge of bibliometric data. This interaction may have increased due to the large-scale availability of bibliometric data and indicators.

The strength of peer review is also its weakness. In the context of national research assessments, the literature identifies the inevitable selectivity of post-publication peer review as a possible problem of exclusively peer review-based evaluation. This may be an area where publication and bibliometric data may add value.

The literature does not currently support the idea of replacing peer review by bibliometrics in the REF. First of all, the existence of strong correlations between bibliometrics and peer review is a necessary but not sufficient condition for replacing peer review by bibliometrics. It depends on which parts of the role of peer review in the REF one wishes to prioritise. Only if the exclusive goal of the REF were to be the distribution of research funding among the universities, could one consider replacing the REF by advanced bibliometric analysis for a number of fields. (However, one should then also consider renaming the funding stream.) Second, the studies that confirmed strong correlations still showed strong variation of ranking results at the level of research institutions. At higher levels of aggregation, the correlation generally becomes stronger. Third, not all fields show strong correlations between bibliometrics and peer review data.

The literature does support the idea of supplementing peer review by bibliometrics (informed peer review). Currently, this concept has not yet been formally operationalised. Bibliometric data may counter specific weaknesses of peer review (its selectivity, particular forms of bias, etc.). Experiments with forms of informed peer review therefore seem the best way forward. Bibliometrics may also help to open up the process of disciplinary peer review to include criteria for societal impact of research.

3. Alternative metrics for research evaluation

The need to evaluate the contributions of researchers, research groups, departments or collections of papers occurs in many situations, including job applications, promotion decisions, research assessment exercises, research funding programme assessments, and grant applications. Although peer judgments are commonly used in such cases, quantitative indicators may sometimes aid the decision making: ‘to inform, but not to determine, judgments of research quality’ (Warner, 2000, p. 453). These quantitative indicators have mainly been based on citations in traditional citation indexes, such as WoS and Scopus.

Although there seems to be agreement that traditional citation databases are the best sources of such data to help peer review and research evaluation, some aspects of intellectual impact may not be well reflected in conventional citation indexes. For example, citation indexes are not comprehensive and are mainly restricted to English language refereed journal articles, with weaker coverage of books and conference papers. Another problem is that many publications may be frequently used during the research process or for other academically related activities, such as teaching, without being formally cited. More generally, citation databases are unlikely to be useful to track the wider impacts of research, such as on business, government and society. Hence, research that has important societal or cultural impacts may be systematically undervalued if assessed with the aid of citation-based indicators. Thus, it is clear that other data sources are needed if quantitative indicators are to be used to aid the evaluation of the wider impacts of academic research.

Peer review seems to be more reliable than citation counting for research evaluation and hence is the first choice in most cases, although subjective perceptions of research quality may cause many different types of bias (Lee et al., 2013) and finding expert reviewers can also be difficult (Weller, 2001). Expert judgments could also be time-consuming and expensive, especially for large research assessment exercises. For instance, in the 2008 UK RAE for biological sciences ‘each panel member assessed an average of just under 1,000 papers within a few months’ (Eyre-Walker & Stoletzki, 2013, p. 7). The number of books per reviewer was also up to 100 in some SSH fields, such as history, in the 2008 UK RAE (Kousha et al., 2011). Hence, it seems that not all submitted research can be reviewed in depth for large-scale research evaluations (Taylor & Walker, 2009; see also: Weller, 2001). Moreover, in the 2001-2003 Italian national research assessment exercise over 6,600 experts (about 22% from overseas) were involved at a cost of about 3.5 million Euros (Franceschet & Costantini, 2011, p. 275). The operating expenditure for the 2008 UK RAE was about £12 million, compared to £5.1 million in 2001 and £3 million in 1996 (http://www.rae.ac.uk/pubs/2009/manager/manager.pdf, p.45).

More than a decade ago, the rise of new ways for scholars to write, communicate and publish research via electronic media (e.g. Kling & McKim, 1999, 2000) led to calls for novel indicators for electronic scholarly communication (Ingwersen, 1998; Cronin, 2001a; Borgman & Furner, 2002). In response, alternative indicators have been developed to capture evidence of types of impact. These *alternative metrics* include *web citations* in digitised scholarly documents (e.g. eprints, books, science blogs or clinical guidelines) or, more recently, *altmetrics* derived from social media sources (e.g. social bookmarks, comments, ratings, microblog posts). Scholars nowadays may also produce and use non-refereed academic outputs, such as multimedia products, datasets and software. It is important to estimate the impact of these non-standard outputs too, if possible, and new usage-based indicators would be needed for this.

In summary, alternative metrics may be helpful when evaluators, funders or even national research assessments need to know 'all kinds of social, economic and cultural benefits and impacts beyond academia' (REF, 2011, p. 4) as well as non-standard impacts inside academia. This literature review provides an overview of the findings of research into many different types of alternative metrics. It also discusses the use of associated indicators for assessing the impact of articles, books and other academic outputs (e.g. science videos, datasets and software). Moreover, summary guidelines are given of the potential advantages and limitations of these alternative metrics over traditional bibliometric indicators.

This document contains no suggestions that alternative metrics (or any other bibliometric indicators) can be used as *replacements* for peer judgments of the quality of individual research outputs. When used, their role should be to *inform* peer judgments, such as by providing a second opinion (perhaps causing reassessments when indicators disagree with peer judgments), by giving additional evidence for marginal cases, by arbitrating when experts cannot resolve disagreements, for a sanity or bias check on overall sets of results (e.g. at the departmental level), or by supplying quantitative evidence to support individual claims for impact by researchers. Nevertheless, indicators might be used as a primary data source, if appropriately constructed and validated, to compare between reasonably large sets of outputs when fine-grained assessments are not needed. For example, funding organisations' research programme evaluations may find some alternative metrics to be useful to indicate whether research funded by one programme tends to have more impact than research funded by another.

3.1. Web-based open access scholarly databases

The web now contains a range of websites hosting free general scholarly databases, such as GS and Google Books (GB), as well as institutional and subject repositories (e.g. ADS, AgEcon, arXiv, CiteSeer, Dryad, PhilPapers, PubMed, RePEc, SSRN – see also http://www.opendoar.org/) some of which form new sources of citation or usage data. These inherit many of the strengths and limitations of traditional bibliometric databases, but with some important differences. This section covers some key databases and GB is discussed in the section on indicators for books.

3.1.1. Limitations of traditional citation databases: A brief overview

The citation-based indicators used for research evaluation are imperfect and have many limitations that should be considered when they are used (see MacRoberts & MacRoberts, 1989, 1996; Moed, 2005a). In October, 2014, Thomson Reuters claimed that its citation indexes cover about 12,000 core journals, 160,000 conference proceedings and 50,000 editorially selected books in science, social science and, arts and humanities (http://wokinfo.com/products_tools/multidisciplinary/webofscience). The growth rate of its Science Citation Index (SCI), however, seems to be slower than that of other comparable databases, suggesting that it may be covering a decreasing proportion of the scientific literature, especially in the social sciences (Larsen & Von Ins, 2010). Moreover, duplicate articles published in high impact journals seem to attract twice as many citations as identical versions published in lower impact journals, suggesting that citations may partly reflect the prestige of the publishing journal (Larivière & Gingras, 2010), or that the audience for an article may be partly a reflection of the place in which it is published. Journal impact factors are particularly controversial and are not recommended for research evaluation purposes because of their many limitations, such as variability over time and the unfairness of comparing between different types of journals and between journals in different fields (Seglen, 1997; Sombatsompop & Markpin, 2005). Citation indictors also cannot be used for recently published papers because they need time to accrue enough citations for a reasonable assessment and in some subject areas, such as the social sciences, research takes longer to be cited (Glänzel & Schoepflin, 1995; Glänzel et al., 2003). Hence, citation indicators should not be used to compare articles in different fields (unless field normalised), published in different years (unless time normalised) or of different types.

Using traditional citation databases for research evaluation in the social sciences, arts and humanities is more problematic than in science and medicine. This is because scholars in these areas are more likely to publish types of publications, and in languages other than English, that are under-represented in, or absent from, traditional citation indexes (Moed, 2005a; Nederhof, 2006; Huang & Chang, 2008). For example, in journalism and library science 4%, in architecture 6%, in the arts 9% and in education 10% of Australian universities' academic publications 1999-2001 were in Thomson Scientific’s citation indexes (Butler, 2008). Similarly, social sciences and arts and humanities fields in the 2001 UK RAE were less well covered in WoS than was science: in law 24% and in art and design about 30% of submissions to the 2001 RAE were in WoS (Mahdi, D'Este, & Neely, 2008). Just under half (48%) of 4,600 publications by researchers from three UK business schools over the period 2001-2007 in business and management were in WoS, whereas GS searches found 66% of these publications, including 90% of the journal articles (Mingers & Lipitakis, 2010). Furthermore, no significant correlations have been found in nine out of 28 subject areas for the 2001 UK RAE between

WoS citations and RAE peer review scores in SSH fields (e.g. education, sociology, history, politics, international studies), whereas in most science fields the correlations were moderate to high (Mahdi, D'Este, & Neely, 2008, p. 16). One of the reasons for the low WoS coverage of humanities RAE submissions is that about 16.5% of all submissions to the 2008 UK RAE were books (monographs, edited books and book chapters), and this was much higher in SSH (31%) than in science (1%) (Kousha et al., 2011). These studies, combined with other evidence (e.g. Hicks, 1999; Archambault et al., 2006; Nederhof, 2006; Huang & Chang, 2008), suggest that the coverage of WoS and Scopus for both articles and books could be insufficient for bibliometric analyses of SSH research, despite the recent inclusion of some books and monographs in both databases.

3.1.2. GS

GS (http://scholar.google.com) is a free online academic search engine that uses automated software to extract citations from the digital publications that it finds online or that are provided by publishers. Researchers not only use GS to search for academic publications (Nicholas et al., 2009; Herrera, 2011), but also to publicise their publications or impact by generating Google Scholar Citations (GSC) profiles (Ortega & Aguillo, 2012). For instance, a survey of 220 science and engineering scientists at one American university showed that GS was second (64.5%) to WoS (66.8%) for routine literature searches out of 18 databases (Hightower & Caldwell, 2010). Similarly, a survey of over 3,000 university faculty in the United States found that Google and GS were the third most ‘often’ or ‘occasionally’ used (about 70%) to find academic publications (Schonfeld & Housewright, 2010).

Although GS was not primarily developed to rival conventional citation indexes, many studies have now compared it against them for research assessment (see Appendix A). GS covers a wider range of academic journals and millions of other scholarly-related publications in different languages and countries, making it particularly worth investigating for impact assessment in areas that are not well covered by WoS or Scopus.

Google does not allow routine automatic gathering data from GS but has made an exception for the *Publish or Perish* software, developed to compute research impact indicators (e.g. the h-index) from GS data (Harzing & Van der Wal, 2008). The superior coverage of GS in computer science and informatics in the UK REF has led to it being recognised as helpful to assist peer review ‘where outputs have been cited extensively outside the body of publications indexed in Scopus’ (REF, 2012, p. 72). A team of scientometricians has also recommended GS citations for the individual assessment of researchers in the EU (after checking), when evaluators, research committees and funders need complementary or wider impact indicators (ACUMEN Portfolio, 2014).

3.1.2.1. GS coverage vs. conventional citation indexes

Regardless of the quality of GS-indexed publications (see next section), GS appears to cover about 88% (100 out of 114 million) of the English-language scholarly documents accessible on the web (Khabsa & Giles, 2014) which seems to be about double the number of WoS scientific records (about 51 million including conference proceedings by September 2014[6]). GS also seems to have comparable coverage of high impact scientific journals. A study conducted in 2006, for instance, found that GS covers 86%, 88% and 81% of the journals in the Thomson Reuters Science, Social Sciences and Arts and Humanities Citation Indexes (Mayr & Walter, 2007). Since then, the current and retroactive coverage of GS appears to have expanded in many science fields (e.g. Chen, 2010; Harzing, 2014, 2013b; De Winter et al., 2014; Orduña-Malea & Delgado López-Cózar, 2014). Many studies have confirmed that GS has greater coverage of international and non-traditional publications (see Appendix A), suggesting that it could be useful for assessing citation impact outside that covered by conventional citation indexes (e.g. Meho & Yang, 2007; Bar-Ilan, 2008b; Kulkarni et al., 2009; Franceschet, 2010a; Kousha et al., 2011; De Groote & Raszewski, 2012; Minasny et al., 2013).

The wide coverage of GS is not universal, however, and its coverage of publishers and other sources varies across fields. For instance, in chemistry the median citation counts of accepted papers derived from WoS, Scopus and Chemical Abstracts (23, 23 and 25, respectively) have been much higher than the GS citation counts (median 1) (Bornmann et al., 2009) and more WoS unique citations were found than GS citations (450 vs. 61, respectively) for 276 sampled chemistry articles (Kousha & Thelwall, 2008b). In contrast, a comparison of the h-index for 5,283 computer scientists derived from GS and WoS showed that the mean h-index from GS (3.54) was higher than the mean h-index from WoS (2.19), but this was not the case for 1,354 physicists (GS h-index mean 6.7 and WoS h-index mean 7.15) (Henzinger et al., 2009). A comparison of citations to 1,000 books submitted to the 2008 UK RAE across seven book-based disciplines also found that both the numbers and medians of GS citations to books were three times as high as the comparable Scopus citations (Kousha et al., 2011). This suggests that, in addition to computer science, GS may be more useful for arts and humanities research than is WoS. The same may be true for business because a study of the publications of Canadian business school faculty members found that the mean number of publications (22), citations (271), and the h-index (4.6) derived from GS were much higher than from WoS (5, 51 and 1.9, respectively) (Amara & Landry, 2012) and the GS mean citations per paper were almost double those of WoS for the research outputs of three UK business and management schools (Mingers & Lipitakis,

---

6. For the number of WoS records the query used in the "publication name" field was: (A* OR B* OR C* OR D* OR E* OR F* OR G* OR H* OR I* OR J* OR K* OR L* OR M* OR N* OR O* OR P* OR Q* OR R* OR S* OR T* OR U* OR V* OR W* OR X* OR Y* OR Z* OR 0* OR 1* OR 2* OR 3* OR 4* OR 5* OR 6* OR 7* OR 8* OR 9*)

2010). Nevertheless, the coverage of GS may have expanded since some or all of these studies were completed.

3.1.2.2. Problems with GS for research evaluation

Despite the substantial, albeit occasionally patchy, GS coverage of publications and citations in comparison with conventional citation indexes, GS data should not be used without extra checking for the evaluation of individuals for a number of reasons. First, GS does not provide transparent information about its indexed sources and its coverage may change substantially over time without warning or notice. Most importantly, however, GS has no clear quality control over its indexed publications. Thus, manipulation of citation counts, automatically generated or deliberately faked documents and references as well as misidentification of authors, publication titles and years are serious concerns for those wishing to use raw statistics from GS for evaluative purposes (e.g. Norris & Oppenheim, 2007; Falagas et al., 2008; Jacsó, 2006, 2008a, 2010 and 2011; Beel & Gipp, 2010a, 2010b; Labbé & Labbé, 2013; López-Cózar et al., 2014). This seems to be particularly problematic for assessments of individual academics (Jacsó, 2008b, 2008c) or articles.

Summary: Consulting GS to locate citations for the impact assessment of research could be helpful when evaluators need a database with wider coverage than that of WoS or Scopus, such as for computer science, probably business, arts and humanities and perhaps many more, although GS does not seem to provide improved coverage in some fields (this may have changed since the research reviewed above). GS also seem likely to be useful for assessments including a substantial amount of non-English documents and perhaps also when recently published or in press publications must be assessed. However, due to a lack of quality control over its indexing of web publications, GS raw data is susceptible to spamming to an extent that it should not be used unfiltered for serious research evaluation purposes. GSC might be useful for the citation statistics in authors' profiles (Ortega & Aguillo, 2014), but not all authors have profiles, and there are problems with citation manipulation and errors in citation attributions.

3.1.3. Patents and Google Patents

A patent is a set of legal rights to an invention within a particular country or set of countries that is usually registered in patent offices for a period of time. Patents contain citations and, intuitively, a citation from a patent indicates that the cited document may have some commercial value or may have helped to generate commercial value. There are differences and similarities between patent and paper citations (for reviews see Meyer, 2000a; Oppenheim, 2000). For patents, both authors and examiners decide which publications should be cited. In fact, patent examiners may add or remove applicant citations based upon judgments of relevance. Thus, patent citations could reflect the citation motivations of both examiners and applicants. Moreover, 'there are many differences between academic and patent citations, which makes it very difficult simply to transfer one theoretical framework from one field to the other. However, both forms of citations have so much in common that findings in one field can be used as inspiration for research in the other' (Meyer, 2000a, p. 111). Assuming that citations from patents can be used as evidence of commercial impacts of research, such as influence on emerging technologies and innovations (Meyer, 2000b; Meyer, 2001; Meyer, 2002), patent databases could be used to research monitoring.

3.1.3.1. Google Patents for locating impact evidence in patents

Google Patents (GP) claims to cover the full text of patents and patent applications originating from the United States Patent and Trademark Office (USPTO) from 1790 and the European Patent Office (EPO) from 1978 (https://support.google.com/faqs/answer/2539193). Hence, whilst not internationally comprehensive, it covers two important sources. The GP full-text search capability makes it possible to locate citations to academic publications within a large number of digitised patents. For instance, a conference paper *Viz3D: Effective exploratory visualization of large multidimensional data sets* by Artero et al. had not received any citations in WoS citation indexes (including conferences) by October 2014. At this date, however, it had been formally cited in 14 patents indexed by GP, suggesting that it may be a type of research that is more useful for inventors than for academics. The number of citations to publications from patents has been previously recommended as one way for academics to demonstrate evidence of the commercial relevance of their research (ACUMEN Portfolio, 2014, p. 42).

Summary: GP citation searches may help to identify some types of commercially relevant research. Nevertheless, its value seems to be patchy in science because patents are not used in many areas of industry, so the lack of a citation from a patent is very far from evidence that an article has had no direct commercial value. Moreover, automatic citation counting from GP is not possible, making it difficult to use in practice for substantial collections of publications in a large-scale research assessment exercise.

3.1.4. Usage indicators from scholarly databases

Usage data are a logical choice to supplement citation counts and digital readership information can be easily and routinely collected, apart from for paper copies of articles. Several early studies have shown that more cited journals tended to be more read (e.g. Stankus & Rice, 1982; Tsay, 1998) and so readership *may* reflect a similar type of impact to that of citations.

Statistics about downloads or views of electronic articles can, in theory, be extracted from local library log files, digital libraries, aggregator services and scientific publishers. In addition, partial usage statistics can be extracted from some social bookmarking tools (Haustein & Siebenlist, 2011). Indicators from this data are based on the assumption that a view or download of a scholarly source tends to indicate someone who has 'an interest or need with regard to a particular resource', although views and full-text downloads of a publication may reflect different degrees of interest or need (Kurtz & Bollen, 2010, p. 6).

Although an early study found no connection between online views and citations for journals (Darmoni et al., 2000), later investigations have found positive associations between citations and downloads, suggesting that the two tend to reflect overlapping types of impact (e.g. Kurtz et al., 2005; Brody et al., 2006; Duy & Vaughan, 2006). A detailed study of astrophysics articles found a strong association between the number of electronic accesses of and the number of citations of online articles based on data from NASA Astrophysics Data System. Although citation counts could predict electronic accesses and vice versa, 'the combination of the two measures of use substantially improves the capabilities of bibliometric measurement' (Kurtz et al., 2005, p. 128). A significant positive Spearman correlation (0.22) has also been found between download rates and citation counts to 1,190 articles published in the journal *Tetrahedron Letters*, during the two years after publication (Moed, 2005b). This correlation increased for downloads made after three months from the publication date (0.35). Similar correlations have also been found for articles deposited to arXiv.org in physics (0.462), mathematics (0.347), astrophysics (0.477), and condensed matter (0.330) and after six months download rates were a good predictor of citation impact after two years (Brody et al., 2006).

Local usage data (e.g. institutional) can also be used for download indicators and one study found them to correlate significantly (r=0.935, 0.624 and 0.681) with the local citation data of researchers for three publishers and one Canadian university. Nevertheless, there was no association between the JIF and journal usage data, suggesting that local citations better reflect journal use than do global impact factors (Duy & Vaughan, 2006). Significant correlations have also been found between local online journal use provided by publishers and local journal citations for 639 journals at the California Institute of Technology (McDonald, 2007). This association was stronger than correlations between local print use and local citations for a set of 458 journals, indicating that online journal data captures more usage than does its print counterpart. Chu and Krichel (2007) examined the relationships between citation indicators (Social Science Citation Index (SSCI) and GS) and download rates for 200 top downloaded papers from the RePEc e-print archive in economics, finding moderate statistically significant correlations between download rates for papers with citations from SSCI citations (0.54) and GS (0.61). Another study also found significant positive correlations between different total impact factor and journal usage factor indicators in computer science, economics and finance, oncology and arts and humanities (except for psychology), although there were some disciplinary differences in the relationships between the citation and download indicators (Gorraiz, Gumpenberger, & Schlögl, 2013).

It is possible to some extent to guess at the audience of an article from the IP addresses of its downloaders. For example, if the IP addresses are all associated with universities then the audience is presumably academic but if a substantial fraction comes from commercial sector organisations then this suggests a wider audience. Although it is technically possible to make such breakdowns and the wider impact evidence that they might generate would be useful for research evaluation, it is difficult to make them robust in practice. Hence, with a few exceptions (e.g. Duin et al., 2012), this approach has not been used for evaluations.

A range of usage-oriented metrics have been proposed that are analogous to classical journal citation indicators such as the 'usage impact factor' (Bollen & Van de Sompel, 2008), 'usage immediacy index' or 'download immediacy index' (Rowlands & Nicholas, 2007; Wan et al., 2010) and 'usage half-life' (Rowlands and Nicholas, 2007; Schloegl & Gorraiz, 2010). There have also been important initiatives to develop platforms to collect and process usage data from publishers, including MESUR (Bollen et al., 2008) and SERUM (Gorraiz & Gumpenberger, 2010), in addition to the COUNTER initiative (www.projectcounter.org) to standardise counting across publishers. Nevertheless, usage statistics are not generally available for research assessment yet and they seem to be relatively easy to spam to some extent.

Summary: Bibliometric indicators do not show the usage of a published work by non-authors, such as students, some academics, and non-academic users who do not usually publish but may read scholarly publications. Usage-based statistics for scientific publications may therefore help to give a better understanding of the usage patterns of documents and can be more recent than bibliometric indicators. Many studies have found correlations between usage and bibliometric indicators for articles and usage data could be extracted from different sources such as publishers, aggregator services, digital libraries and academic social websites. Nonetheless, the usage statistics could be inflated or manipulated and some articles may be downloaded or printed but not read or may be read offline or via different websites such as authors' CVs and digital repositories (Thelwall, 2012). Hence, integrated usage statistics from different sources such as publishers' websites, repositories and academic social websites, if they are not manipulated in advance, would be optimal for global usage data. This does not seem to be practical yet, however.

3.2. Citations and links from the general web

It is possible to extract information from the web in order to identify citations to publications, hence using the web as a huge and uncontrolled de-facto citation database. This data collection can be automated, such as through the Bing Applications Programming Interface (API), making the web a practical source of this type of citation data. The free software Webometric Analyst (http://linkanalysis.wlv.ac.uk) can run automatic searches through the Bing API for this purpose.

3.2.1. Link analysis

Over a decade ago webometric researchers attempted to assess online impact by counting web hyperlinks on the basis that, like citations, they were inter-document connections that may tend to confer authority on their targets (Almind & Ingwersen, 1997; Rousseau, 1997). This is also the idea behind Google's PageRank algorithm and so is an intuitively credible idea. It led to the 'Web Impact Factor' (Ingwersen, 1998), which was similar to the Journal Impact Factor but based on hyperlinks and being applicable to any collection of websites. Online mentions of academics' names (Cronin et al., 1998) have also been proposed as a method to identify the wider impacts or fame of academics. These initiatives all examined whether web-extracted metrics could provide data for impact assessment that could extend traditional citation indicators (Cronin, 2001a). Many other early investigations also exploited analogies between web links and citations to develop indicators for the impact of journal websites or online articles (Harter & Ford, 2000; Smith, 1999; Vaughan & Hysen, 2002; Vaughan & Thelwall, 2003). On a larger scale, studies of sets of university websites revealed that link counts correlated with the amount of research produced by universities, as measured by the RAE or similar exercises (e.g. Thelwall, 2001; Smith & Thelwall, 2002; Thelwall & Harries, 2003). Nevertheless, the removal of hyperlink search facilities from all major commercial search engines has undermined the use of link data for the web impact assessment of research, although alternative methods have been suggested, including 'URL citations'(Kousha & Thelwall, 2006) and 'linked title mentions' (Sud & Thelwall, 2014b), as discussed below.

Summary: Early studies found that counts of web hyperlinks to online articles, journal websites and university websites correlated with traditional citation metrics or other indicators of research productivity or impact. Although link-based metrics have not been used to assess the research of individuals, the number of links to a university website (external inlinks) is one of the indicators used for measuring the visibility of academic institutions (Aguillo et al., 2006) in the Webometrics Ranking of World Universities (http://www.webometrics.info). Hyperlink counts are now less easy to obtain and are probably not useful for assessing the impact of individual papers, academics or even research groups but may be helpful as a visibility indicator at the entire institution level, although link spam is widespread and hyperlinks can be generated automatically in large numbers for legitimate reasons, such as to connect related online databases or Wikis.

3.2.2. Web and URL citations

Vaughan and Shaw (2003, 2005) used the term 'web citation' to refer to a mention of an exact article title in a web page, proposing counts of these as a new potential impact indicator and showing that they tended to correlate with traditional citation-based metrics. Web citations, in this sense, can easily be identified by searches for article titles in commercial search engines. These web citation searches may return matches in the reference lists or text of any type of document on the web. In contrast, a 'URL citation' is a mention the URL of an online scholarly work (e.g. an open access article) in a web page. Both web and URL citations can be gathered manually from the online interfaces of commercial search engines or automatically by submitting queries to Bing through its API, although any more than 5,000 queries per month will need to be paid for.

URL citation counts have been used as an alternative to web citation counts with similarly promising evidence that they correlate with traditional citation counts (Kousha & Thelwall, 2006, 2007a). URL citations have the advantage that, unlike article titles, they are normally unique and hence unambiguous, but the disadvantage that many citations of online publications omit the paper's URL or use a DOI (Digital Object Identifier) as an indirect pointer. Moreover, previous studies have shown that general web or URL citation searching with commercial search engines gives results that need extensive manual checking to identify online citations in formal research publications because most web or URL citations seem to be created for non-scientific reasons, such as (arguably) library reading lists and online copies of journal tables of contents. For instance, out of 854 web citations to 46 library and information science journals, only 30% were citations from other publications (Vaughan & Shaw, 2003) and only a quarter of online citations to journal articles in biology, physics, chemistry, and computing represent citation impact from references in other web documents (Kousha & Thelwall, 2007b). URL citations are probably less useful now than when they were originally conceived because of the use of complex URLs in some modern publishers' websites and the rise of DOIs as an alternative method for pointing to online documents.

Summary: Web or URL citations to publications can be located by commercial search engines (Google manually and Bing automatically) from almost any type of online document, including blog posts, presentations, clinical guidelines, technical reports or document files (e.g. .pdf files) and there is evidence (although not recent) that they can be indicators of research impact. In theory, then, web and URL citations could be used to gather evidence about the scholarly impact of research if they are filtered to remove non-scholarly sources. In contrast, unfiltered web or URL citation counts are easy to spam and many citations are created for navigation, self-publicity or current awareness and so it does not seem likely that they would genuinely reflect the wider impacts of research, without time-consuming manual filtering out of irrelevant sources.

3.3. Citations from specific parts of the general web

In addition to searching for citations from the general web, citations can be counted from specific parts of the web, including types of website and types of document. This information can be extracted from appropriate searches in commercial search engines and automated via the Bing API. The discussions below cover online presentations, syllabi and science blogs, although there is also some evidence that mentions in news websites and discussion forums may also be useful (Costas et al., 2014; Thelwall et al., 2013). Citations from online grey literature seem to be an additional useful source of evidence of the wider impact of research (Wilkinson et al., 2014), but there do not seem to be any systematic studies of these.

3.3.1. Online presentations

Conferences are important for sharing scientific results in some areas of science (Drott, 1995). In computer science and engineering, refereed conference papers are particularly important research outputs. For example, over 40% of citations to highly cited publications in computer science are from proceedings papers (Bar-Ilan, 2010). The share of cited proceedings in Thomson Scientific citation indexes 1980-2005 was about 20% in computer science, about 13% in electrical engineering and electronics, and 11% in civil engineering. Proceedings papers tend to receive citations earlier than does the cited literature in general (Lisée et al., 2008).

Conference papers are presumably initially given with the aid of presentation files (e.g. in Microsoft .ppt and .pptx or Apple .key). Presentations in the same format may also be used for teaching and informal seminars. These presentations may then be posted online and become searchable by commercial search engines or available through slide-sharing sites such as slideshare.net or slideshow.com. This gives them the potential to be used for a new type of online citation analysis.

Although most scientific results in presentations will be formally published later in proceedings or journals, some academic presentations may never appear elsewhere. For instance, there are about 11,000 citations to PowerPoint presentation files (.ppt and .pptx) in the references of Scopus publications, a quarter of them in computer science (authors' data, see http://www.koosha.tripod.com/citationtopowepoints.jpg), suggesting that their content was useful enough to be cited by other research even though they were not formally published.

3.3.1.1. Citations from academic presentations

Citations from academic publications can be systematically gathered by automatically submitting queries to commercial search engines, such as through the Bing API, using bibliographic information for the query and specifying presentation files only in the results (e.g. adding filetype:ppt to each query). Based on a study of about 1,800 WoS-indexed journals in 10 science and 10 social science fields, citations from online presentations are not numerous enough for general impact assessment, but presentation citations could be helpful to identify important articles in popular magazines like *Scientific American* and *Harvard Business Review* (Thelwall & Kousha, 2008). A classification of reasons for mentioning social science journals in 756 PowerPoint files from American university websites found that about 60% occurred in formally cited references and 15% were in course reading lists, indicating that the majority (about 75%) represented a type of intellectual impact. However, about 15% of the journals were mentioned for reasons not reflecting intellectual impact, such as CVs and publishers' lists of journals (Thelwall & Kousha, 2008). Presentations are easy to spam, however, even if they are only searched for in academic websites.

Summary: Citations from online presentations can be automatically collected through web searches and could perhaps be a helpful source of impact in conference-based fields (e.g. computer science and engineering), although they seem to be too rare for this data to be worth routinely collecting for research assessment purposes. Moreover, they are easy to spam and so should not be used for important evaluations without manual checks.

3.3.2. Online course syllabi for educational impact

Course syllabi often record the most important textbooks for students to read and so are a logical source of information about whether books and articles are useful in teaching. There have been many content analyses and comparative studies of the contents or structure of academic course syllabi (e.g. Pieterse et al., 2009; Mishra, Day, Littles, & Vandewalker, 2011; Homa et al., 2013), but they have not been used for research assessments. Nevertheless, the educational impact of publications seems to be important for teaching-based fields, and particularly in the less hierarchical knowledge structures of SSH, where textbooks, books and monographs can have educational value rather than, or in addition to, research impact (e.g. Gurung & Martin, 2011; Gurung et al., 2012).

Mentions of publications (e.g. textbooks or articles) in online academic course syllabi can be automatically retrieved from the web using appropriate Bing API searches, making syllabus mentions a practical indicator for research assessment. One study searched for mentions of over 70,000 journal articles published in 2003 in online course syllabi in multiple fields, finding substantial numbers of mentions in some social science disciplines (e.g. political sciences and information science), but syllabus mentions were less than 13% as frequent as citations in each of the fields analysed and were less than 0.1% as frequent in mathematics. A case study of library and information science articles showed that the articles that were most recommended in academic syllabi tended to be reasonably highly cited but that the converse was not true (Kousha & Thelwall, 2008a). This confirms that some articles can have more educational influence than research impact and since this study more syllabi or course reading lists may be available online, especially from an international perspective, perhaps allowing more inclusive teaching impact assessment.

Summary: Statistics about the uptake of academic publications in academic syllabi may be useful in teaching-oriented and book-based fields, where the main scholarly outputs of teaching staff are articles or monographs for which students are an important part of the audience, or textbooks. It is practical to harvest such data from the minority of syllabi that have been published online in the open web and indexed by search engines, but it seems that such syllabus mentions may be useful primarily to identify publications with a particularly high educational impact rather than for the systematic assessment of the educational impact of research. Syllabus mentions have most potential for the humanities and social sciences, where they are most common and where educational impact may be most important.

3.3.3. Science blogs

Blogs are continuing to be very popular. For instance, the number of posts and comments in blogs hosted by WordPress.com has apparently increased from 15 and 13 million in November 2011 to 44 and 58 million in August 2014, respectively (https://wordpress.com/activity/). Today, there are also many science blog hosting services, such as scienceblogs.com, blogs.nature.com and blogs.plos.org, where academics can discuss scientific issues. For instance, ResearchBlogging.org includes 'blog posts about serious peer-reviewed research, instead of just news reports and press releases' (http://researchblogging.org/static/index/page/about).

The important contribution that academic blogs can make to informal scholarly communication has been widely recognised and analysed (e.g. Ewins, 2005, Luzón, 2007, 2009; Davies & Merchant, 2007; Kirkup, 2010; Shema et al., 2012; Mewburn & Thomson, 2013). In terms of composition, about 60% of a sample of 126 ResearchBlogging.org bloggers were affiliated with academic institutions, 65% were graduate students, and 72% of ResearchBlogging.org blogs were written by one or two male authors, indicating important gender differences (Shema et al., 2012). Academics appear to be less dominant in one German scientific blogging platform, however, but 60% declared that dissemination of their field of research to the general public was their main reason for blogging (Puschmann & Mahrt, 2012). This is consistent with blogs being an alternative platform to present ideas and 'to write outside the boundaries of traditional academic publication' (Davies & Merchant, 2007, p. 177). In terms of attempting to generate wider impact, however, there has been a claim that academics can use blogs to make their scholarly works or ideas accessible to readers outside academia and receive feedback from the public, and 'such a function contrasts with conventional modes of academic performance premised on expertise and mastery' (Gregg, 2006 [online]).

Several qualitative investigations have focused on motivations for blogging science. Kirkup (2010) interviewed six academic bloggers in different fields and positions. She argued that academic blogging is 'an emerging academic practice, and a new genre of scholarly writing, which could become an important activity and skill for a professional academic' (Kirkup, 2010, p. 82). Another interview-based study with 11 academic bloggers in three European countries concluded that science blogs can function as a platform to disseminate and publicise scientific contents to other potential readers, to express opinions in a manner that is rarely possible in other academic writing, to keep up-to-date, and to interact with other researchers (Kjellberg, 2010). In contrast to academic articles which are primarily written for other scholars in the field, academic blogs can potentially establish different relationships with audiences through informal writing styles, non-standard citation practices and social networking (Mortensen & Walker, 2002). Nevertheless, blogs can provide inputs to formal scholarly communication because citations to major blogs (e.g. ScienceBlogs and BlogSpot) from Scopus publications have increased from 21 citations before 2003 to just under 5,000 in 2011 (Kousha & Thelwall, 2014b). In support of this, a content analysis of 100 academic blogs found that academic cultural critique (41%) and research dissemination (40%) were the two most common types of contents and the intended audience of the majority of blogs were academics and researchers (Mewburn & Thomson, 2013). Nevertheless, using Researchblogging.org and a combination of webometric and bibliometric techniques and mapping, a study of a set of 295 chemistry blog posts about peer-reviewed research, found that scientific discourse on the web is more immediate and contextually relevant than the traditional academic literature and focuses on non-technical implications of science (Groth & Gurney, 2010).

3.3.3.1. Science blogs as a source of intellectual impact

Blog posts may include links or references to other publications and these citations could perhaps be gathered to form an indicator of the impact of the cited research. Hyperlinks in academic blogs seem to be created for many informal scholarly reasons, such as to increase the visibility and collaboration of bloggers in their scientific community or to publicise their research outputs (Luzón, 2009). One study found that about 30% of academics frequently linked to articles, newspapers and other documents that they discussed or provided commentary about (Luzón, 2009). It is possible to manually search for blog citations with Google Blog Search to try to gather evidence about the social impact of the cited research, although one study found that few articles from two information science journals were cited in blogs (citation means: 0.34 and 0.44) in comparison to WoS (11 and 8) (Kousha et al., 2010b).

Another study using blog citation data for 13,300 medical and biological sciences articles from altmetric.com (Adie & Roe, 2013) found them to correlate with WoS citation counts at a low but statistically significant level ($r=0.201$; $p<0.01$) (Thelwall et al., 2013). Another investigation found that for 58% and 68% of journals published in 2009 and 2010, respectively, articles blogged in ResearchBlogging.org tended to subsequently receive more citations than did other articles from the same journal (Shema, Bar-Ilan, & Thelwall, 2014). Both of these studies concluded that academic blog citations could be an alternative source for the impact assessment of research.

Summary: Research may be cited and discussed in blogs by academics or non-academics in order to debate with or inform other academics or a wider audience. Blog citations can perhaps be considered as evidence of a combination of academic interest and a potential wider social interest, even if the bloggers themselves tend to be academics. In addition, the evidence that more blogged articles are likely to receive more formal citations shows that blog citations could be used for early impact evidence. Nevertheless, blog citations are not straightforward to collect and may need to be provided by specialist altmetric software or organisations, and are easy to spam.

3.3.4. Other sources of online impact

In addition to the types of web citations discussed above, preliminary research is evaluating online clinical guidelines, government documents and encyclopaedias. Online clinical guidelines (for a discussion of issues see Manchikanti et al., 2012) could be useful for medical research funders to help them to assess the societal impact of individual studies (Kryl et al., 2012). In support of this, one study extracted 6,128 cited references from 327 documents produced by the National Institute of Health and Clinical Excellence (NICE) in the UK, finding articles cited in guidelines tend to be more highly cited than comparable articles (Thelwall & Maflahi, in press).

With over 4.6 million articles in English (http://en.wikipedia.org/wiki/Wikipedia:Size_of_Wikipedia), many of which have references, Wikipedia may reflect wider uses of research. In support of this, one study found significant correlations between WoS citations to scientific journals and citations from Wikipedia (Nielsen, 2007). However, another study on a sample of over 24,000 articles published by the Public Library of Science showed that only 5% were cited in Wikipedia, whereas 80% had at least one Mendeley bookmark (Priem et al., 2012). Thus, Wikipedia citations seem to be too rare for routine use in research evaluation, even though they could be automatically extracted without too much difficulty from copies of Wikipedia freely provided by its owners (http://en.wikipedia.org/wiki/Wikipedia:Database_download).

3.4. Citations, links, downloads and recommendations from social websites

A number of readership-based indicators have been proposed, with umbrella terms like 'reading factors' (Darmoni et al., 2000), 'readership rates' (Kurtz, et al., 2000 and 2005) and 'alternative metrics' (Bollen et al., 2005). Even before this library usage statistics, such as photocopy requests (Cooper & McGregor, 1994), or journal re-shelving counts (Tsay, 1998) had been proposed as alternatives to bibliometric indictors. The advent of the social web, however, has seen an explosion in both the range of indicators that could be calculated as well as the ease with which relevant data can be collected (even in comparison to web impact metrics). Of particular interest are comments, ratings, social bookmarks, and microblogging (e.g. Taraborelli, 2008; Neylon & Wu, 2009; Priem & Hemminger, 2010). Although there have been many concerns about validity and the quality of altmetric indicators due to the ease with which they can be manipulated (Birkholz & Wang, 2011; Rasmussen & Andersen, 2013), there has been interest in evaluating them from the scientometrics community (e.g. Wang et al., 2013). Elsevier (via Scopus), Springer, BioMed Central and *Nature* have all added altmetrics to articles in their collections.

Although the term 'altmetrics' refers to indicators for research assessment derived from the social web (Priem et al., 2010), some scholars have proposed other names, such as 'influmetrics' (Cronin & Weaver, 1995; Rousseau & Fred, 2013), 'metrics of social impact' (Eysenbach, 2011), 'Scientometrics 2.0' (Priem & Hemminger, 2010), or just 'non-standard indicators' (Donovan & Butler, 2007; Mohammadi & Thelwall, 2013). The term 'alternative metrics' seems to be gaining currency now, however, as a catch-all for web-based metrics, perhaps because of the existence of a company, altmetrics.com, with the term altmetrics prominently in its name.

A range of altmetrics have been shown to correlate significantly and positively with bibliometric indicators for individual articles (e.g. Priem, Piwowar, & Hemminger, 2012; Thelwall et al., 2013; Costas et al., 2014), giving evidence that, despite the uncontrolled nature of the social web, altmetrics may be related to scholarly activities in some way. This is perhaps most evident when the altmetrics are aggregated to entire journals (Alhoori & Furuta, 2014; Haustein & Siebenlist, 2011) rather than to individual articles, however.

Social usage impact can be extracted from a range of social websites that allow users to upload, or register information about, academic publications, such as Mendeley.com, Twitter, Academia.edu, and ResearchGate.net. These sites can be used for assessing an aspect of the usage of publications based on numbers of downloads, views or registered readers. Although all of these sources are partial and biased, the lack of global usage-based statistics (see the download and views subsection, 3.4.1) means that social media citations might be useful in some contexts for research impact assessment. Note, however, that the exact value of some of these impact statistics can depend partly on the data source providing the figures rather than just on the social website analysed (Jobmann et al., 2014)

3.4.1. Faculty of 1000 web recommendations

Scientific papers are typically peer reviewed before being published in journals or, sometimes, conference proceedings, although biases in peer review are acknowledged to exist (for an in-depth discussion see Weller, 2001). Several countries (e.g. UK, Australia and Italy) also employ expert post-publication peer judgments of (normally) peer reviewed research in order to allocate public funds, although this essentially double peer review approach has been criticised on the basis that the first should be sufficient (Bence & Oppenheim, 2004). In contrast, there have been arguments that post-publication reviews can be used as a further quality control mechanism for the critical analysis of a published work by a wider spectrum of experts (e.g. Crotty, 2012; Hunter, 2012; da Silva, 2013). Informal pre-submission reviews may also have also been common in some science fields, such as physics, for unrefereed manuscript drafts in eprint archives (e.g. arXiv).

Faculty of 1000 (F1000) is a commercial post-publication peer review website that recommends and rates selected biomedical science publications. In October 2014, it claimed to gather the judgments of

5,000 academics and experts in biology and medicine to review journal articles that they have read. An early investigation found a medium significant correlation (Spearman r= 0.445) between F1000 ratings and peer judgments from Wellcome Trust (a UK-based funding research institution) for a small sample of 48 original research papers (Allen et al., 2009). However, articles that were highly rated by experts were not always the most highly cited, and vice versa. A study of 1,530 articles published in seven leading ecological journals in 2005 compared citations from WoS with F1000 recommendations, finding that the 103 articles recommended by F1000 tended to attract more citations (median: 23) than did typical publications in the dataset (median: 16) (Wardle, 2010). There were outliers, however, because 11 highly cited articles (cited 120-497 times) were not recommended and just under half (46%) of the recommended publications were not highly cited. Geographical biases in reviewers and uneven coverage of F1000 could be the reasons why F1000 was unable to identify all high impact articles. F1000 should not be used in areas for which it has only partial coverage, however, such as ecology (Wardle, 2010).

A study of 1,397 journal articles published in 2008 in *Genomics and Genetics* found statistically significant correlations, albeit low, between F1000 judge rating scores and WoS/Scopus/GS citation counts (0.295, 0.293 and 0.290, respectively) and between F1000 scores and Journal Impact Factors (r=0.359) (Li & Thelwall, 2012). F1000 ratings "are useful for post-publication evaluation purposes. However, the lower correlations suggest that they measure different perspectives of research" although "It may not be feasible to have the same system for other disciplines" because of the greater importance of biomedical research (Li & Thelwall, 2012, p.549). Another study compared F1000 scores and Scopus citations to 344 and 533 medical science articles published in 2007 and 2008, respectively, finding low but statistically significant Spearman correlations for both years (r=0.383 and r=0.300, respectively). A lower correlation was found between the number of labels assigned by F1000 reviewers and Scopus citation counts (r=0.201) (Mohammadi & Thelwall, 2013). The study suggested that F1000 labels and the judge ratings could be useful in research evaluation exercises when the importance of practical findings needs to be recognised because citation counts do not always reflect applied value of research. This aligns with the findings discussed above for references in clinical guidelines.

Another systematic study of F1000 ratings assessed correlations between seven bibliometric indictors from Thomson Reuters InCites and F1000 article scores. Of 5,204 papers from InCites in cell biology or immunology published in 2008, 125 (2.4%) had F1000 ratings (Bornmann & Leydesdorff, 2013). The 'Journal Actual/Expected Citations' indicator explained only 1% of the variance in FFa (the lowest correlation), whereas 'Percentile in Subject Area' explained 20% of the variance in FFa scores (the highest correlation), suggesting that F1000 scores tend to reflect something substantially different from, albeit overlapping with, citation counts.

The above results align with another study that also reported low but significant correlations between F1000 article scores and JIFs (r=0.28). However, a further analysis using standardised regression coefficients between assessor scores and JIFs and numbers of citations showed that F1000 ratings are more strongly dependent on the JIFs than on the number of citations, suggesting that post-publication assessors ‘might tend to rate papers in high IF journals more highly irrespective of their intrinsic merit’ (Eyre-Walker & Stoletzki, 2013, p. 2). Alternatively, the journal in which an article is published may be a better indication of its overall value than its citation count.

The largest-scale academic investigation of F1000 so far matched all 132,662 F1000 recommendations with WoS, finding 95,385 (93%) matching publications in WoS that were recommended 124,320 times in F1000 (Waltman & Costas, 2014). However, only about 2% of the biological and medical sciences publications had F1000 recommendations, showing that its coverage is perhaps too low for systematic research assessment exercises. They also found that about half of the recommended articles were published in the top 10% most highly cited journals, although three quarters of the top 1% most highly cited articles had not been recommended (Waltman & Costas, 2014), aligning with the above-mentioned study (Wardle, 2010). In terms of associations between F1000 scores and citation indicators, the number of F1000 recommendations that publications received significantly correlated with their citations (Pearson r=0.26) and Journal Citation Scores (Pearson r=0.34). Similar low correlations were found between both F1000 maximum recommendation scores and weighted numbers of recommendations with citation counts and Journal Citation Scores. The low correlations confirm that recommendations and citations probably reflect different types of impact to some extent.

Summary: Post-publishing peer review indicators could be very helpful for research assessment and F1000 seems to be particularly valuable for its expert ratings of articles that may reflect values that do not necessarily attract citations, such as utility for clinical practice. It is not clear yet, however, whether the ratings are biased by perceptions of JIFs. Moreover, the coverage of F1000 in biomedical science is very low (perhaps about 2%), and it is not clear that any similar system would be financially viable or as useful for any other research fields.

3.4.2. Mendeley and other online reference managers

One method to capture publication usage evidence from social media tools is to count bookmarks in online reference management software, such as Mendeley (Henning & Reichelt, 2008), CiteULike (Bogers & Bosch, 2008), Zotero (Ritterbusha, 2007), Bibsonomy (Borrego & Fry, 2012), and Connotea (Hull et al., 2008). These websites allow users to register for free and then enter information about publications of interest. The websites then help users to create reference lists from their saved publication information and share their libraries of reference information with others. The assumption behind counting users bookmarking a publication is that they are likely to use the articles for their research, and perhaps cite them later, or use them in other academic activities (teaching or lectures). This is supported by evidence from a survey of Mendeley users (Mohammadi, 2014). After individual users have bookmarked articles in Mendeley the number of bookmarks for each article in the system can be automatically downloaded with the Mendeley API and exploited as usage information. Mendeley seems to be the most attractive tool for altmetric data because it relatively easy to automatically extract bookmark counts from the Mendeley API, its data seems to be high quality (see below), and it seems to have at least as many users as other reference managers (see Li et al., 2012). Mendeley bookmarks positively and moderately correlate with counts of citations to published journal articles in many different research fields, as discussed in detail below (Bar-Ilan, 2012; Bar-Ilan et al., 2012; Li & Thelwall, 2012; Li et al., 2012; Mohammadi & Thelwall, 2014; Thelwall et al., 2013; Zahedi et al., 2014).

An early investigation of papers published in *Nature* (793) and *Science* (820) in 2007 found significant Spearman correlations between bookmarks and WoS citation counts (0.559 and 0.540) and GS citations (0.592 and 0.603) for both journals (Li et al., 2012). There were also significant, but lower (0.304, 0.396) correlations between CiteULike bookmarks and citation counts, perhaps because 93% of the articles had at least one bookmark in Mendeley in comparison to 60% for CiteULike. A larger study of 1,397 journal articles from F1000 in genomics and genetics found strong and statistically significant correlations between Mendeley bookmarks and WoS/Scopus/GS citations (0.686, 0.682 and 0.694, respectively) which were larger than the correlations between CiteUlike bookmarks and citation counts (0.354, 0.346 and 0.377, respectively) (Li & Thelwall, 2012). About 30% of WoS articles published in 2008 in engineering, chemistry and physics had at least one Mendeley bookmark, in comparison to 60% for clinical medicine (Mohammadi et al., in press). The correlations between citations and Mendeley bookmarks were moderate and higher in clinical medicine (r=0.463) than in chemistry (r=.369), engineering and technology (r=.327) and physics (r=.308). Based upon a sample of over 24,000 articles in seven journals from the open-access publisher Public Library of Science (PLoS), 80% had at least one Mendeley bookmark, in comparison to 30% in CiteULike. For articles published in *PLoS ONE*, *PLoS Biology* and *PLoS Pathogens* there were moderate Spearman correlations (0.3, 0.4 and 0.4 respectively) between Mendeley bookmarks and WoS citations (Priem et al., 2012). Similarly, for 1,706 PLoS Biology research articles (published up to May 20, 2013) 95% and 65% had Mendeley and CiteULike bookmarks (Fenner, 2013). Hence, Mendeley's coverage of sciences seems to be generally high, and bookmark counts seem to correlate moderately with citation counts.

In SSH, Mendeley's coverage is lower than in the sciences. The largest-scale investigation of Mendeley so far analysed WoS articles published in 2008 in five social sciences (n=62,647) and in five humanities (n=14,640) disciplines, finding low and medium Spearman correlations between Mendeley bookmarks and citation counts (Mohammadi & Thelwall, 2014). About 58% and 28% of the WoS articles were in the Mendeley catalogue in SSH, respectively, suggesting that Mendeley's coverage of the academic literature may not be as high outside of science as was found for science in previous studies. The correlation was higher in Business and Economics (0.573), Information Science and Library Science (0.535) and Psychology (0.514) than in Religion (0.363), Philosophy (0.366) and Literature (0.403), perhaps because citations are less common, less important and used for different purposes in the humanities. Moderate correlations (WoS 0.458, Scopus 0.502, GS 0.519) and extensive Mendeley coverage (97%) have been found for articles published in the *Journal of the American Society for Information Science and Technology (JASIST)* during 2001-2011 (Bar-Ilan, 2012), probably because of its disciplinary focus on libraries and information management. Edited books and monographs are important in the humanities and Mendeley has limited coverage of them (see below). Based on a sample of 310 journal articles in 2012 in humanities from 30 Swedish universities, Mendeley had greater coverage (61%) than other altmetric data sources and on average articles had 3.4 readers in Mendeley compared with 2.4 GS citations (Hammarfelt, 2014), and so Mendeley's coverage of the humanities, whilst low, may still be higher than comparable sources.

The results above are consistent with the findings of a large random sample of 20,000 WoS-indexed publications across different fields 2005-2011, which found that 63% had at least one Mendeley bookmark and this was a bit higher for articles (66%) (Zahedi et al., 2014). The overall Spearman correlation between Mendeley bookmarks and WoS citations was moderate (r=0.49).

A different approach compared indicators derived from arXiv, Scopus and Mendeley for publications from a sample of 100 European astrophysicists (Bar-Ilan, 2014). The Mendeley readership counts were much lower than the Scopus citations (e.g. 90 readers compared with 1,168 Scopus citations) and the overlap between Scopus and Mendeley was about 22%. In contrast with previous studies, a much lower Spearman correlation was found between Scopus citations and Mendeley readership counts for articles (r=0.227), perhaps because many physicists do not use Mendeley.

Mendeley bookmark counts may not be frequent enough to aid the impact assessment of scientific books. For example, out of 2,739 scientific monographs indexed by the Book Citation Index (BKCI) in 2008, only 7% had at least one Mendeley bookmark and in science and medicine there were very low correlations between Mendeley bookmarks and BKCI and GB citations (Spearman r=0.106 and 0.139 respectively, n=718) (Kousha & Thelwall, in press). Similarly, only 7% of 54 English books published by Swedish universities were in Mendeley (Hammarfelt, 2014). Other alternative metrics for wider book impact assessment are discussed below.

Some of the above results may underestimate the value of Mendeley through using incomplete methods to identify the number of readers for articles. The best method currently seems to be to combine DOI searches with traditional queries (Zahedi et al., 2014).

Mendeley records some information about users and reports this with bookmark counts in the form of a breakdown of the three most common types of user for the publication in several categories. This shows, for example, the majority countries of origin and disciplines of readers. One of these categories, ‘Other professionals’, could be used, in theory, to identify non-academic users of research but this category is rare, suggesting that Mendeley is mainly used inside academia (Mohammadi et al., in press).

An important property of Mendeley is timeliness: it seems likely that Mendeley bookmarks will appear before citations because citing authors would presumably bookmark referenced articles in Mendeley before completing their research and submitting an article or book for publication. Hence a Mendeley bookmark should appear about a year before the citation is indexed. In support of this, there is some evidence of the value of Mendeley readers as early citation indicators in one field (Maflahi & Thelwall, in press).

Summary: Mendeley readership bookmarks seem to be the most promising altmetric indicator because of the ease of automatic data collection, the wide coverage of articles (a majority of recent articles are bookmarked in Mendeley in most fields checked) and evidence of low, moderate and strong correlations between readership bookmarks and citation counts. Moreover, Mendeley may give earlier evidence of impact than can citation counts. Nevertheless, Mendeley is not subject to quality control, could be spammed by asking other Mendeley users to bookmark articles or to create fake Mendeley profiles, and does not seem to be used much for books. In addition, Mendeley seems to reflect a similar kind of impact to that of citation counts (rather than reflecting educational impact or other wider research impacts) and so it is not clear that it would be useful additional information to supplement citation counts, except perhaps for indications of early impact for recently-published articles. Mendeley can also provide information about readers of publications in terms of their fields, countries and academic positions, which may be useful for detailed evaluations of funding programmes.

3.4.3. Twitter and microblog citations

Twitter is one of the most popular web social network and microblogging services, allowing free short instant posts of up to 140 characters. There are apparently about 500 million tweets per day (http://www.internetlivestats.com/twitter-statistics), making Twitter a huge source of comments about many topics. A study of tweet citations to PubMed articles from altmetric.com found that they were more numerous than 10 other social media outputs, including Facebook wall posts, Google+ posts and blog citations (Thelwall et al., 2013) and another did the same for articles tweeted with an DOI or other identifiable ID July-December, 2011, finding Twitter again to have the wider coverage (13% of a multidisciplinary sample of WoS articles) than the other altmetrics considered (Costas et al., 2014). Hence, tweets are particularly promising from a purely numerical point of view. These studies did not compare Twitter against Mendeley, however, and the comprehensiveness of altmetric.com's data is unknown.

Several studies of academic-related Twitter use have surveyed researchers (e.g. Letierce, Passant, Decker, & Breslin, 2010; Letierce, Passant, Breslin, & Decker, 2010) or analysed the content of tweets sent during conferences or meetings (e.g. Ross et al., 2011; Desai et al., 2012; McKendrick et al., 2012; Hawkins et al., 2014; Neill et al., 2014; Wen et al., 2014. See also Weller et al., 2011). In brief, these studies indicated that Twitter is used to share basic information about conference talks, discussions and academic papers. There are disciplinary differences in how researchers use Twitter, however. For instance, conversations in tweets in one small study were more common in digital humanities and cognitive science (both 38%), astrophysics (31%) and history of science than in biochemistry and economics (both 16%). In biochemistry, 42% of tweets are retweets, whereas in nine other fields the proportion varied from 18% in social network analysis to 33% in sociology (Holmberg & Thelwall, 2014). Whilst these tweets could theoretically be read by any Twitter user, the posting of such content by academics is not evidence of successfully attracting a wider audience for research.

Assuming that counts of tweets sent by researchers or academics that mention a scholarly work may be an indication of intellectual impact of the tweeted publications, several investigations have examined tweets as a potential source of altmetrics. Priem and Costello (2010) interviewed 28 academics and coded 2,300 tweets with hyperlinks from them, finding that about 6% were Twitter citations. Whilst half of the Twitter citations directly cited a resource, articles were also cited indirectly, such as via discussions. More promisingly, however, tweet citations were much faster to appear than conventional citations, with 40% occurring within a week of publication.

Eysenbach (2011) compared over 1,570 tweets with links to 55 articles published in his *Journal of Medical Internet Research* (2009-2010) against subsequent citation counts from Scopus and GS 17 to 29 months later. Highly tweeted papers were 11 times more likely to be highly cited than their less-tweeted counterparts and tweets correlated with later citations moderately well. A study of 4,600 articles submitted to the arXiv.org preprints archive during a half-year period also found significant moderate correlations between Twitter mentions with article citations (Pearson r=0.452) and arXiv downloads (Pearson r=0.505) and Twitter mentions had shorter delays than did arXiv downloads for predicting citations (Shuai et al., 2012). This relatively high figure is probably misleading because the Pearson correlation is sensitive to skewed data, such as citation and tweet counts. For example, removing the top two tweeted articles reduced the correlation by 0.2 (Shuai et al., 2012). In contrast, another study found negative low Spearman correlations (Spearman r=-0.236) between tweets and citations to a set of PubMed articles from 2010 (Thelwall et al., 2013; see also: Haustein et al., 2014; Costas et al., 2014). This was found to be due to more recent articles having been tweeted more frequently whereas older articles had been cited more. This time the effect was strong enough within individual years to create a negative correlation. Using more sensitive statistical methods designed to correct for all time biases, however, the same data was shown to contain a positive association between tweets and citations (Thelwall et al., 2013). Another analysis of altmetric data, this time with articles tweeted with a DOI or other identifiable ID July-December 2011, found low positive Spearman correlations between tweet counts and the total number and field normalised number of citations of publications (Spearman 0.167 and 0.141, respectively) (Costas et al., 2014). This low positive correlation, in contrast to the small negative correlation for the previous study with similar data, is probably due to the time span being half of a year rather than a year. Overall, however, it is clear that whilst tweets associate with citations, this association is very weak and is only evident for units of analysis below a year or specially designed non-correlation statistical measures.

A content analysis of 270 tweets linking to articles in four journals (*PLOS ONE*, *PNAS*, *Science*, and *Nature*), four digital libraries (Wiley, ScienceDirect, Springer, and JSTOR) and two DOI URLs attempted to identify why articles were tweeted and whether there was evidence of uptake outside academia. The results found no evidence of this, with 83% of the tweets merely repeating an article title or a brief summary of it without giving any context that could be evidence of the type of impact that the articles had had. Only 4% of tweets were positive about the articles and none were critical, suggesting that tweet links to articles reflect the popularity or visibility of an article rather than a particular type of impact (Thelwall Tsou, Weingart, Holmberg, & Haustein, 2013).

An international limitation of Twitter is that its uptake is not uniform across the globe and so its results will be biased against areas of research that are popular in countries that tend not to use it. For example Twitter seems to be rarely used and sometimes blocked in China, with Sina Weibo being popular instead, and has also been blocked in Iran.

Old tweets must be bought from a data reseller, such as GNIP or DataSift, or from altmetric providers if they have sufficient coverage. Although Twitter gives free automated access to current Tweets through its API, this includes only a fraction of tweets sent and also excludes tweets that are older than a few weeks. GNP or DataSift have bought all tweets and have permission to sell it onwards, and so provide a theoretical solution for the problem of identifying all tweets linking to a given set of articles. Nevertheless, the charges for the data are based on query complexity and volume and this is likely to make tweets prohibitively expensive to buy except perhaps for short periods of time and for small collections of papers with a common URL base. Since altmetric companies may have relevant historical data then they may provide a cheaper source if they have sufficiently comprehensive data.

Summary: Tweet citations appear much more quickly after publication than do conventional citations – days rather than years. Although Twitter is used by a wide section of the public outside academia, no study yet seems to have found evidence of a substantial non-academic audience in Twitter for academic research. Instead, it seems to be mainly used for information-sharing between academics as well as for other types of informal scholarly communication. In general, tweets correlate at a low positive level with citations if they are analysed over time periods of under six months, but this may not be enough for Tweet counts to be a useful indicator, except perhaps for identifying individual articles with very high levels of tweeting.

3.4.4. Facebook and Google+ citations

Facebook wall posts and Google+ posts seem to be similar to tweets in the way that they are used and the findings for Twitter probably also apply to citations and links from them too. Some studies using data from altmetric.com have confirmed that articles tend to be more highly cited if they are mentioned in Facebook or Google+, although they appear to be much less common than are tweets for academic articles (Costas et al., 2014; Thelwall et al., 2013). Facebook data seems to be more difficult to collect systematically, however. A limitation with using general social network sites is that their uptake varies internationally and some countries have their own popular sites, such as VK in Russia and Tencent Qzone in China. Hence any impact data from Facebook and Google+ would be internationally biased.

3.4.5. Academic social network sites: Usage and follower counts

Academics may use a variety of methods to increase the visibility of their research, such as publishing web CVs and maintaining profiles in one or more academic social network sites, including ResearchGate.net and Academia.edu. The latter help scholars to disseminate research and to interact with other academics but also provide some usage and impact-related statistics that may be helpful for impact assessment. However, most scholars probably do not use these sites or use them but not to systematically record their publications. For instance, a survey of 100 researchers in an Indian university showed that under a quarter used ResearchGate to find out about others' research (Chakraborty, 2012) and an investigation of 1,500 highly cited scientists working at European institutions revealed that few had profiles in major social network sites (e.g. a quarter had LinkedIn profiles and even fewer had Academia profiles) (Mas Bleda et al., 2014).

3.4.5.1. ResearchGate

ResearchGate.net is a free social network site for academics, researchers and students that claims over a million members (http://www.researchgate.net/about). Each member can report information about themselves and upload or list their publications, whether peer-reviewed or not. Its uptake is not comprehensive, however. For example, out of over 2,090 teaching or research staff in Nicolaus Copernicus University in Poland, about 14% had ResearchGate profiles (Stachowiak, 2014). For registered publications, ResearchGate provides the number of full-text downloads, views and citations (based on information in its database). It also provides some information for individual members, such as the total number of publication views and downloads, as well as how many followers they have (Kadriu, 2013). Rankings of institutions based on ResearchGate statistics correlate moderately well with other rankings of academic institutions (e.g. The Times Higher Education Ranking or The CWTS Leiden Ranking), suggesting that ResearchGate use broadly reflects traditional academic capital at the institutional level (Thelwall & Kousha, 2014b).

Summary: ResearchGate views, downloads and citation counts would be potentially useful for the assessment of individual articles when authors register on ResearchGate and upload their articles to their profiles, especially prior to formal publication, but these statistics can be easily manipulated or spammed (e.g. usage statistics may be inflated by authors or a machine). Moreover, it is difficult to automatically gather ResearchGate statistics because it does not have an API and it is probably used only by a minority of academics and so it is likely to have weak coverage of the academic literature. In terms of the wider influence of academics, however, ResearchGate's use for academic social networks may be valuable to assess the social impact of scholars within academia based on followers, although there is no evidence yet that this would be effective and the only partial usage of ResearchGate suggests that it might be problematic.

3.4.5.2. Academia.edu

Like ResearchGate, Academia.edu has facilities for sharing information about publications and their full text. Initiated by an Oxford University philosopher, Academia.edu claims over 14 million academic members, over 3.4 million papers and over 15.7 million unique monthly visitors (https://www.academia.edu/about). Academia.edu provides some usage statistics for individual papers and authors (aggregating the results for all of their papers) as well as their numbers of followers. A study of user profiles in philosophy departments found that faculty members tended to attract more profile views than did students but female philosophers did not attract as many profile views as their male counterparts, suggesting that academic capital drives philosophy usage of the site more than friendship and networking (Thelwall & Kousha, 2014a). Conventional bibliometric indicators (h-index and citations) did not correlate significantly with any Academia.edu metrics (profile views and document views) for philosophers, perhaps because more senior academics use the site less extensively or because of the range of informal scholarly activities that cannot be measured by bibliometric methods. Hence it is not clear whether Academia.edu could provide useful indicators to help in evaluations of individual scholars, and no evidence has been gathered yet to evaluate the value of Academia.edu usage statistics for individual articles.

The top 15 broad research interests registered by Academia.edu users are related to the humanities and social sciences (excluding computer science in third), indicating that it is heavily used by academics in these fields and suggesting that its greatest potential is outside science (Thelwall & Kousha, 2014a, p. 731).

Summary: Usage statistics from Academia.edu seem to have the same potentials and spam limitations as those from ResearchGate, especially perhaps in the humanities, where bibliometric indicators probably do not reflect the usage of research by students or other academics who do not usually publish journal articles. Nevertheless, there is little hard evidence as to the value of the indicators that can be derived from its data and, like ResearchGate, it does not have an API and therefore data collection is not simple.

3.5. Indicators for book impact assessment

Research evaluation in book-oriented fields is more challenging than for article-based subject areas because counts of citations *from* articles, which dominate traditional citation indexes, seem insufficient to assess the impact of books. The BKCI within WoS is a recent response to this issue (previously noted in Garfield, 1996a) since journal citations on their own might miss about half of the citations to books (Hicks, 1999). Some academic books are primarily written for teaching (e.g. textbooks) or cultural purposes (e.g. novels and poetry) and citation counts of any kind may be wholly inappropriate for these.

Books were more frequent in SSH (31%) than in science (1%) in the 2008 UK RAE, and many of these books (art, music and literary works) may have merits that are not reflected by conventional bibliometric methods. Moreover, the main sources of citations to humanities books are other books (Thompson, 2002; Kousha & Thelwall, 2014a). Even today, the Thomson Reuters BKCI and Scopus index a relatively small number of books (50,000[7] and 40,000[8] as of September 2014, respectively) and this may cause problems for bibliometric analyses of books (e.g. Gorraiz et al., 2013; Torres-Salinas et al., 2012, 2013). Expert peer judgment of books seems to be by far the best method but it is even more time-consuming and expensive than article peer assessment because books tend to be longer and some aspects of book impact (e.g. teaching or cultural) could be particularly subjective (see Weller, 2001). In response, different alternative sources have been investigated for book impact assessment, including syllabus mentions, library holding counts, book reviews and publisher prestige.

Many of the indicators discussed elsewhere in this review can also be used for books but have not yet been evaluated for this purpose. Since books seem to be read offline, download indicators are probably not relevant, however.

---

[7] http://wokinfo.com/products_tools/multidisciplinary/webofscience

[8] http://blog.scopus.com/posts/scopus-content-book-expansion-project-update

3.5.1. Google Books

GB (http://book.google.com) contains a large number of academic and non-academic books based upon digitising the collections of over 40 libraries around the world as well as partnerships with publishers (http://books.google.com/intl/en/googlebooks/about/). GB seems to cover at least 30 million volumes (Darnton, 2013), although the exact figure has not been disclosed. Several studies have shown that the coverage of GB is quite comprehensive, however. For instance, about 84% of 401 randomly selected books from WorldCat (a global catalogue of library collections) in different languages (Chen, 2012) and about 80% of 1,500 Hawaiian and Pacific books from a university library were found in GB (Weiss & James, 2013b). A study of 400 English and 400 Spanish language books from a university library also found that almost all English (92%) and Spanish (89%) titles were in GB, suggesting small language differences in comprehensiveness (Weiss & James, 2013a). A study of 2,500 pages from 50 randomly selected books found that less than 1% had legibility errors (James, 2010) and so GB seems to be a fairly comprehensive and good quality source of digital books. Nevertheless, due to copyright considerations, GB does not always reveal to users the full text of the books that it has indexed.

3.5.1.1. Google Books citations for impact assessment

Although GB is not a citation index and provides no citation statistics of any kind, it is possible to manually search it for academic publications and hence identify citations to these publications from digitised books (Kousha & Thelwall, 2009; Kousha et al., 2011). GB could be useful because citations from books have been largely been invisible in traditional citation indexes and the current book citation search facilities in Scopus and WoS cover relatively few books that are predominantly in English and from a small number of publishers, which is problematic for citation impact assessment in book-based disciplines (Gorraiz, Purnell, & Glänzel, 2013; Torres-Salinas et al., 2012, 2013).

Several studies have explored the potential use of GB citations for research assessment. A comparison of citations from GB searches with WoS citations to 3,573 journal articles in 10 SSH and science fields found GB citations to be 31%-212% as numerous as WoS citations in SSH, but only 3%-5% as numerous in the sciences checked, except for computing (46%) (Kousha & Thelwall, 2009). There were significant positive correlations between GB and WoS citation counts for all fields, although they were higher in computer science (.709), philosophy (.654) and linguistics (.612) and lower in chemistry (.345) and physics (.152). Despite GB not being a citation index, its citation search capability clearly has promise as an additional source for the citation impact of research. A follow up study manually searched and compared citations from GB with citations from Scopus (cited references search in articles) to 1,000 books submitted to the 2008 UK RAE in seven book-based fields (Kousha et al., 2011). Overall, GB citations were 1.4 times more numerous than Scopus citations. In history, the median number of GB citations (11.5) was higher than for both GS (seven) and Scopus (four) citations. Moreover, in communication studies and law the median number of GB citations (11.5 and six, respectively) was roughly three times as large as the Scopus citations (four and two, respectively). There were also high, significant and positive correlations between GB and Scopus citation counts in all fields (ranging from 0.616 in law to 0.833 in sociology). Thus, in many humanities subject areas citations from books to books may be more substantial than citations from journal or conference papers to books and hence GB seems to be valuable, and perhaps in some cases the most valuable source, for the impact assessment of books. This was confirmed with a study of citations to 14,500 monographs in the Thomson Reuters BKCI against GB automatic searches in 24 subject areas because GB citations were 103% to 137% higher than BKCI citations (including journals) in the humanities, except for tourism (72%) and linguistics (91%), 46% to 85% in the social sciences, but only 8% to 53% in the sciences. There were also moderate correlations between the GB and BKCI citation counts in SSH, suggesting that citations from GB and BKCI could reflect different aspects of impact with most BKCI citations coming from WoS-indexed journals rather than books (Kousha & Thelwall, 2014a). Good results have also been obtained from GB for counts of citations to books in a non-English nation, Malaysia (Abrizah & Thelwall, 2014), and it seems that both GB and GS could be very helpful for non-Western countries seeking to assess the impact of their locally-published books and monographs, especially in SSH.

Although GB citation searches can be automated through the GB API with searches constructed from the bibliometric information of books and articles, the raw data needs to be filtered because not all matches are genuine citations. Nevertheless, a highly accurate (over 90%) filtering process has been developed to deal with this issue and so automatic GB searching is practical (Kousha & Thelwall, 2014a). However, for the individual assessment of academics extra manual checking might be necessary, and citations to documents with titles or authors containing non-ASCII characters may be less reliable.

No study so far has compared GB citations with expert judgments and so it is not clear for which fields high GB citation counts would be accepted as evidence of the value of monographs or book chapters. For example, some humanities fields may argue that scholars may be valued for the uniqueness of their expert knowledge, such as in Old Norse, and the more unique the knowledge, the fewer the citations a work may expect to get, irrespective of its quality. Similarly, mathematicians may argue that the significance of pure mathematics is not reflected in its uptake but in the complexity of the problem solved. Citations for maths would then be irrelevant and may even have an inverse relationship with significance if few others are capable of understanding particularly complex contributions. Engineers may also claim that citations are irrelevant because they tend to reflect the extent to which a contribution is theoretical rather than applied, whereas the community may particularly value applied contributions that lead to engineering projects without necessarily being cited. A related case is that in some humanities, such as area studies or classics, the ability to publish in languages other than English may be valued, even though it would be effectively penalised by any type of citation counting due to the dominance of English as a scholarly publishing language.

Summary: GB citation search seems useful for assessing academic research impact of publications, especially in book-oriented fields, but only for subjects for which experts agree that GB citations (or citations in general) tend to reflect a desired property of research. GB citations provide unique and more numerous citations from books in comparison to conventional citation databases in many arts and humanities fields and some social sciences, but not in the sciences. GB citation counts may tend to reflect the teaching or cultural impact of books (e.g. textbooks or novels), when they are cited in other contexts than research, such as for educational or literary reasons. In contrast to GS, which indexes web publications, GB indexes published books and hence seems less likely to be spammed, although it is possible to publish fake or artificial books through cheap publishers (e.g. search GB for Jesse Russell, Ronald Cohn) and this could be used to generate self-citations. Moreover, although it is possible to use automatic GB citation searching with a high level of accuracy in terms of the results (90%), this level of accuracy is probably lower than for the major current citation indexes.

3.5.2. Libcitations

For a long time librarians have used statistics such as the demand for photocopies of publications to assess the usage or impact of their library collections (e.g. Cooper & McGregor, 1994). For example, library journal use (counting journal re-shelving) has been found to correlate with the citation counts and impact factors of 835 medical journals in one general hospital library (Tsay, 1998). Although the original purpose of library usage statistics was as a tool for collection management they may also be useful for research evaluation, particularly for books for which electronic download statistics are not available. A much more recent study compared library loan statistics for the most borrowed monographs from two European university libraries (Granada and Vienna) with citation counts (WoS and GS). Loans and citations did not significantly correlate, however, except for GS citations for textbooks or manuals from the Vienna sample (Cabezas-Clavijo et al., 2013). There does not seem to be an initiative to systematically collate any such usage data from libraries, however, and so it is not currently a practical option.

Another straightforward way to assess the impact of a book is to assess its sales or to count how many libraries have bought it. White, Boell, Yu et al. (2009) coined the term “libcitation” for the number of libraries holding a book, as calculated from national or international union catalogues, and suggested that this may give an indication of the cultural benefit of books from the social sciences and humanities. A comparison of the libcitations of books from several Australian academic departments in history, philosophy, and political science, concluded that libcitation statistics can potentially “allow the departments to be compared for cultural impact” (White et al. 2009, p. 1083).

Significant correlations have been found between library holdings and WoS citation counts for books produced by the Faculty of Humanities at Leiden University (Pearson's r= 0.29). The correlation was higher for books in English (r=0.39), but insignificant for books in Dutch, perhaps because libraries outside the Netherlands and Flanders may be reluctant to stock Dutch books and scholars internationally may be reluctant to read and cite them, and so there may be less data for such books (Linmans, 2010). A much larger-scale study compared Scopus citations to 59,000 history books and 42,000 literature books referenced in Scopus-indexed journals with library holding counts from the Association of Research Libraries (ARL), non-ARL libraries and all libraries. Low Spearman correlations were found, ranging from 0.288 for citations and ARL library holdings to 0.244 for citations and non-ARL libraries. The low but significant relationships confirm that 'citations and "libcitations" [] measure (partially) different dimensions' (Zuccala & Guns, 2013, p. 359). Finally, a comparison of WorldCat library holdings with citations from Thomson Reuters BKCI and GB to 2,739 academic monographs from 2008 also found significant but low positive correlations in the social sciences (r=0.145 for BKCI and 0.234 for GB, n=759), arts and humanities (r=0.141 for BKCI and 0.268 for GB, n=1,262). However, in science the correlation was only significant between library holdings and GB citations (0.112, n=718) (Kousha & Thelwall, in press).

It is also possible to gather and collate library holding information from a defined set of libraries, if universal coverage is not wanted (Torres-Salinas & Moed, 2009).

Summary: National or international library holdings statistics can indicate library interest in books and seem to reflect a different type of impact to that of citations, perhaps including educational and cultural impacts. These statistics are relatively simple to collect automatically from the Online Computer Library Center (OCLC) WorldCat library holding catalogue (http://www.worldcat.org) with more than 2.2 billion items from over 72,000 libraries in 170 countries (http://oclc.org/worldcat/catalog.en.html). This data, which is based upon book holdings and hence would be costly to spam, seems promising for assessing the wider influence of books in SSH based on the information needs of users, teaching staff and researchers. Whilst more detailed borrowing statistics might be even more useful, this data does not seem to be currently available.

3.5.3. Online book reviews

Scholarly book reviews are important in some fields and are 'an academic genre with measurable features' (Hartley, 2006, p. 1194). An early investigation reported a high association (r=0.620) between the number of reviews in the Book Review Index and the number of library holdings in the OCLC database for 200 novels (Shaw, 1991), suggesting that book reviews could be a usage or popularity indicator that may reflect wider cultural impacts. Moreover, there is evidence that sociology monographs (n=420) with positive reviews attract considerably more citations (from Social SciSearch) than do monographs with negative reviews (Nicolaisen, 2002), and so the content of a review may be important in an academic context. Nonetheless, the relationship between the number of book reviews and citations could differ between subject areas (Gorraiz et al., 2014).

Online book reviews, such as those at Amazon.com, could theoretically be used to generate indicators for the wider impacts of books based upon feedback from readers inside and outside academia. One study found low but significant Spearman correlations between the numbers of Amazon reviews and citation metrics for 2,739 academic monographs published in 2008 (Kousha & Thelwall, in press). The correlations were higher in the social sciences (0.223 for BKCI and 0.182 for GB, n=759) and arts and humanities (0.189 and 0.188, n=1,262) than in science fields (0.121 and 0.158, n=718), indicating that Amazon book review counts may partially reflect scholarly impact and may reflect wider impacts such as teaching, cultural or social influence. The relatively low correlations are not surprising given the low correlation previously found for library holdings (see above).

Summary: Counts of reviews and sentiments of reviews of academic books seem to both be useful indicators of the reception or wider uptake of scholarly books. Academic book review databases such as *Choice: Current Reviews for Academic Libraries*, with many book reviews and recommendations (7,000 reviews per year) by over 35,000 editors, experts and librarians in the field (http://www.ala.org/acrl/choice/about) could be a useful altmetric source for research evaluation of books, especially in the arts and humanities. Amazon reviews can be automatically extracted, making them a possible choice, but are easily spammed and so should not be used for formal evaluations.

3.5.3.1. Book review sentiments

Every day many comments are posted on Twitter, YouTube, Facebook, blogs and forums. Sentiment analyses of these social texts are used commercially to assess public opinion about products, services or popular events (e.g. Jansen et al., 2009; Thelwall et al., 2011; Wei et al., 2010; Bai, 2011) and automatic methods are reasonably effective at identifying sentiment in social web texts (e.g. Pang & Lee, 2008; Paltoglou & Thelwall, 2012). The SentiStrength software, for instance, reports sentiment based on a dual scale of 1 (no positive sentiment) to 5 (strong positive sentiment), and -1 (no negative sentiment) to -5 (strong negative sentiment) and is optimised for tweets and other short social web texts (Thelwall et al., 2010; see http://sentistrength.wlv.ac.uk).

Despite the commercial success of sentiment analysis, it has been only rarely assessed for scholarly-related social texts such as comments on research articles. One study, however, estimated sentiments in Amazon.com book reviews about academic monographs and compared them with citation indicators and found significant but low correlations between BKCI and GB citation counts and book reviews sentiments, indicating that monographs with more citations tend to have more positive Amazon.com reviews (Kousha & Thelwall, in press). The Spearman correlations between the positive and negative sentiment strengths of book reviews with BKCI citations were higher in the social sciences (0.216 and -0.218, respectively) and in arts and humanities (0.174 and -0.181) than in science (0.108 and -0.100).

Summary: Comments or reviews on books or other academic outputs web could give evidence of sentiment and this may help to give more accurate results than just counting these comments in the social web. Research into this approach is still at an early stage, however.

3.5.4. Publisher prestige

In the absence of effective citation counts for the impact assessment of books, there have been attempts to use publisher prestige as a simple way to identify more important books. For example, book impact assessment experts in economics in one study believed that 'books should have the same weight as any other publication and should be restricted to those published by major academic presses or a few prestigious commercial publishers' (Donovan & Butler, 2007, p.237). There have been attempts to evaluate the prestige of publishers with surveys for library collection management and research assessment purposes. Metz and Stemmer (1996), for example, surveyed collection development officers in academic libraries about the prestige of different publishers, with university presses being found to be highly regarded. They also mentioned other ways for assessing publishers, such as faculty suggestions, personal reading and winning awards. They believed that subject differences, the existence of specialised publishers and the necessarily subjective nature of judgments were all problems for assessing publishers.

A survey of 603 American political scientists generated rankings of scholarly publishers based upon their publication and reading preferences, with university presses forming the top 10, followed by a mix of university and commercial presses in the top 20 (Garand & Giles, 2011). A much larger survey of Spanish researchers and faculty members with over 3,000 responses has been used to create ranked lists of publishers in SSH, with Oxford University Press, Cambridge University Press and Routledge being the top most prestigious publishers across all of the studied fields, whereas there were large differences in the rankings of the other publishers (Giménez-Toledo et al., 2013). This confirms the existence of disciplinary differences in publisher prestige (see also The Scholarly Publishers Indicators project http://epuc.cchs.csic.es/SPI/).

In contrast to reputational surveys, bibliometric indicators from BKCI have been used to create Book Publishers Citation Reports by analogy with the Journal Citation Reports (Torres-Salinas et al., 2012). This study ranked publishers across 19 SSH fields in terms of the production (e.g. number of books/book chapters indexed) and impact (e.g. total/average citations per book and percentage of non-cited items) of the publishers. The over-representation of English language books, unbalanced coverage of publishers and partial coverage of BKCI were all identified as problems with this approach, however. Citations from Scopus and matching data from WorldCat.org have also been used to rank 50 academic book publishers in history. Oxford University Press, Cambridge University Press and Routledge were again the top three publishers based on total citations and citations per book. Nevertheless, the process of matching, cleaning and standardising bibliographic data of books was difficult, which is a practical limitation (Zuccala et al., in press 2014).

Summary: Reputational surveys, libcitation and citation indicators can help to identify prestigious scholarly publishers. A combination of all of the above may be more useful for rating (rather than ranking) academic publishers of books or monographs as long as other factors, such as geographical, language and disciplinary differences are taken into consideration when they are used. Nevertheless, the construction of publisher prestige indicators may be time consuming and it may be acceptable to allow evaluators to use their own field judgments about the relative prestige of publishers if they do not have to assess books out of their area of expertise.

3.6. Indicators for the impact of non-refereed outputs

Although scholars seem to be evaluated mainly based upon their traditional publications, especially in research assessment exercises, they also produce other outputs that may have substantial value, such as scientific videos, images, datasets and software. Moreover, in some subject areas, non-standard outputs, such as artworks, exhibitions, performances and compositions, may dominate. For some of these there may be plausible indicators, such as audience size, art gallery prestige, composition commissioner prestige, art sales or sales prices. In most cases, however, it is likely that the contributions of individual works are so varied that any data presented to support an impact case would not be directly comparable with other available data, although it could be presented as evidence to support a specific argument about the contribution of a work. This section covers the small minority of potential indicators for non-refereed outputs that have been investigated so far, all of which are online. The lack of research into indicators that are not online partly reflects the relative ease with which they can be gathered but perhaps also the absence of a drive to create indicators for non-refereed arts and humanities outputs.

3.6.1. Scientific data

In some fields, such as genetics, data sharing is vital and datasets are significant research outputs (Borgman, 2012). About 86%, 79% and 64% of the datasets reported in research articles in forensic, evolutionary and medical genetics, respectively, are shared (Anagnostou et al., 2013) and a survey of 1,329 scientific members of the National Science Foundation-funded DataONE project indicated that the majority (85%) were interested in using datasets by other researchers, if they were easily accessible (Tenopir et al., 2011). An international survey of about 370 researchers in the field of biodiversity science showed that 84% agreed that sharing article-related data was a basic responsibility, and only 11% disagreed. Nonetheless, over 60% were unwilling to share primary data before the final publication of their articles (Huang et al., 2012).

Due to the significant role of research datasets in some subject areas, there has been a call for a ‘Data Usage Index’ (DUI) by analogy with conventional citation indexes, such as WoS and Scopus, so that data usage indicators could be developed to recognise the work of the dataset creators (Chavan & Ingwersen, 2009; Ingwersen & Chavan, 2011). Alternative indicators, such views, saves, discussions, and recommendations, have also been suggested for dataset impact assessment (Konkiel, 2013). All of these indicators could help to encourage data sharing by recognising popular datasets and creators. This recognition already occurs indirectly to some extent because a study of 85 cancer microarray clinical trial publications with shared datasets showed that just under half of the trials with publicly available data received about 85% of the aggregate citations and clinical trials with publicly shared data were cited around 70% more frequently than those without (Piwowar, Day, & Fridsma, 2007). Thomson Reuters has recently launched the ‘Data Citation Index’ that ‘indexes a significant number of the world's leading data repositories of critical interest to the scientific community, including over two million data studies and datasets’ (http://wokinfo.com/products_tools/multidisciplinary/dci/) and so dataset citation analysis is likely to become routine and simple when this matures.

Summary: Assessing the impact of academic datasets is important in some fields. This could become possible in the future with the Thomson Reuters Data Citation Index or other alternatives, if they are effective. Whilst most researchers probably do not create datasets, this would allow those that do to claim credit for it and would also encourage data sharing.

3.6.2. Software

In certain fields, such as software engineering and bioinformatics, software can be an important scholarly outcome. Programmers or software engineers may spend time and effort to design and develop useful software and tools for the research community or the public, and free scientific software may be heavily downloaded by researchers or other end users. Some computer programs may also have a significant social, health or educational impacts. For instance, over 400,000 copies of AIDA, a free educational computer program about diabetes (http://www.2aida.net/), have been downloaded and 580,000 simulations have been run on AIDA websites (http://www.2aida.net/aida/logstats.htm).

A range of alternative indicators has been suggested to monitor the success of software projects, such as the number of downloads (e.g. Crowston et al., 2004; Rossi et al., 2010), reuse of programming code, the number of users, and user ratings and satisfaction (Crowston et al., 2003). Alternatively, the online popularity of software could be assessed based on search engines results (Weiss, 2005). It would be useful to have a software citation index to help to reflect the impact of scholarly software in the future. Without this, creators could perhaps choose their own indicator to help demonstrate the value of their work.

Summary: The publication of software seems to be usually overlooked in research evaluations. It would be useful to have a software citation index to help to reflect the impact of scholarly software in the future. Until then, creators could choose their own indicator to help demonstrate the value of their work, although it could be easily spammed.

3.6.3. Science videos

Online scholarly videos are increasingly produced and used by academics for real-time scientific demonstrations, live conferences, presentations, and course lectures. For instance, the *Journal of Number Theory* and the *Journal of Visualized Experiments* have dedicated YouTube Channels for their articles. Over 1,800 Scopus publications have cited at least one YouTube video in their reference lists (as of December 2011) and there has been a constant growth in the citing of online videos from three citations in 2006 to 719 citations in 2011[9]. A content analysis of 551 YouTube videos cited by articles showed that in science and in medical sciences over three quarters of the cited videos had scientific content (e.g. laboratory experiments or academic talks), whereas in the arts and humanities about 80% of the YouTube videos had art, culture or history themes (Kousha et al., 2012). Hence, online videos are a tiny but growing proportion of academic outputs and can even have value for research. Nevertheless, it is hard to quantify the impact of videos even if they are obviously successful (e.g. Haran & Poliakoff, 2012).

A prominent venue through which science and technology information can be communicated to the public is the TED Talks video series. These videos contain curated lectures by academics, artists and others and reach a wide audience. An investigation into TED videos found that few were formally cited in academic publications but a range of metrics including views, comments and comment sentiments were better impact assessment indicators because even academic talks that are highly viewed may not be cited in research (Sugimoto & Thelwall, 2013). For instance, a TED talk video by a social psychology professor, 'Your body language shapes who you are', from June 2012 had been viewed online 20.8 million times but had received only two Scopus citations, suggesting a much greater impact on society than on the scientific community. The metrics in YouTube can be easily spammed but the numbers may be large enough to make effective spamming difficult for this initiative.

Summary: Online videos are used by academics to communicate findings to other scholars or to disseminate findings to (a section of) the public. These videos may take a long time to create and may make important contributions to science, to education or to science communication with the public. Hence, although videos are probably ignored in almost all current research evaluation exercises and

[9] The number of Scopus publications citing at least one YouTube video has dramatically increased to 2,026 in 2013 (see http://www.koosha.tripod.com/youtubecitation.jpg)

videos are awkward to systematically assess because few researchers produce them and they can have very different audiences (from field specialists to the general public) and can be hosted in different ways, it would be valuable to at least allow academics to make the case for the impact of their videos. In this context, usage indicators such as views, comments and comment sentiments would be most appropriate (see Thelwall, Sud, & Vis, 2012), although they are easily spammed.

3.6.4. Academic images

There are now many scientific or artistic images produced by scholars that are on the web and in some cases these are the main outputs of scholarly activities. For instance, specialised photographs of deep astronomical objects are major research outputs in astrophotography (Schröder & Lüthen, 2009). Scientific images also have many applications in different fields, such as in the biological sciences (Glasbey & Horgan, 1995) and for medical diagnoses (Lim, Feng & Cai, 2000). In art and documentary photography the main scholarly outputs are photographs of people, places, or nature. These pictures may appear in publications or be shared online. For instance, *National Geographic* magazine has a worldwide reputation not only for its articles but also for its high quality photographs of wildlife, world culture and scientific advances. It also provides some social media statistics for Facebook likes, Tweets and Google+ for some pictures. More generally, interesting scientific pictures may also be useful for educational and science communication purposes.

There have been attempts to develop metrics as indicators of the type or image usage. For example, tag usage within university image groups in Flickr can be a helpful indicator of social influence (Angus et al., 2008), although not to the same extent as F1000 tags (see above). The number of copies of a science picture on the web may also be an indicator of the level of interest in it, particularly if it is copyright-free. This is possible to count using the TinEye image search engine, as shown by an investigation into academic images from NASA's astronomy picture gallery (Kousha et al., 2010a). Only 1.4% of these pictures seem to have been used in academic publications, but 37% had been used for educational or other scholarly-related reasons, indicating their wider impact.

Summary: Assessing the impact of academic images is important for some academics that produce them and is difficult because images may be used in different ways and for different reasons. However, a combination of text searches (e.g. photographer name, title of image or its URL citations) image searches (e.g. TinEye) and social statistics such as comments, views or tags in Flickr may be useful for their creators as sources of evidence for their uptake, providing that steps are taken to avoid spamming.

3.7. Alternative metrics conclusions and recommendations

As summarised above, there is empirical evidence that a wide range of indicators derived from the web for scholars or their outputs are related to scholarly activities in some way because they correlate positively and significantly with citation counts. In many cases these metrics can also be harvested on a large scale in an automated way with a high degree of accuracy (see Appendix B for methods to obtain alternative metric data). Nevertheless, most are easy to spam (e.g. see Dullaart, 2014) and nearly all are susceptible to spam to some extent. Moreover, whilst a few seem to reflect types of impact that are different from that of traditional citations, none except Google patent citations and clinical guideline citations clearly reflect wider societal impact. In addition, many of them seem to be too rare to help to distinguish between the impacts of typical publications, but could be useful to give evidence of the impact of the small minority of high impact articles. Overall, then, despite the considerable body of mostly positive empirical evidence reviewed above, with some exceptions alternative metrics do not seem to be useful to capture wider social impact and do not seem to be robust enough to be routinely used for evaluations in which it is in the interest of stakeholders to manipulate the results. In other words, alternative metrics do not seem to be suitable as a management tool with any kind of objective to control researchers (Wouters & Costas, 2012). Even if no manipulation took place, which seems unlikely, the results would be suspected to be affected by manipulation and in the worst case the results would be extensively manipulated and scientists would waste their time and money on this manipulation.

In case spamming by academics (which is a completely different type of offence to research fraud, see Steen, 2011) may be thought to be unlikely, the RePEc (Research Papers in Economics) archive managers believe that many authors try to deliberately manipulate views-based or downloads-based public article rankings, despite the lack of direct financial rewards derived from these rankings. For example RePEc abstract views and download statistics 'are subject to manipulation, as one could repeatedly download a paper to increase its count. For this reason, various information about the abstract viewer or downloader are recorded to prevent repeat counts' (Zimmerman, 2013, p. 254). More seriously, 'various checks and balances are implemented to recognize abnormal behaviour, mostly from authors trying to manipulate the statistics. Obviously, these safeguards are not revealed here, but let it be known that a human eye has a final look at the server logs in these cases and that several authors have been caught' (Zimmerman, 2013, p. 254). Hence, RePEc managers apparently believe that transparent automated manipulation detection would be ineffective and that human checking is necessary even with secret manipulation detection algorithms.

Table 4. Summary of alternative metric indicators.

Object	Indicator	Impact type	Evidence of value	Spam	Method of collection
Acad-emics	Academic web site followers	Scholarly	Correlations with h-index	Yes	Manual
Articles, papers	Downloads or views	Scholarly	Correlations with citations	Yes	From publishers?
Books	Google Books Citations	Scholarly; educ.; cultural	Correlations with citations	Yes?	Automatic collection via API
Books	Library holdings = Libcitations	Scholarly; educ.; cultural	Correlations with citations	No?	Automatic collection via API
Books	Reviews, sentiment	Scholarly; public	Correlations with citations	Y (web); N (mag.)	Automatic from web, manual from mags?
Data-sets	WoS citations	Scholarly or public	None	No?	From WoS in the future
Images	Views, copies, tags	Scholarly; educ.; cultural	None	Yes	Automatic collection via API in some cases, TinEye/Google image
All pubs.	Blog citations	Scholarly; educational; public	Correlations with citations	Yes	Commercial altmetric provider
All pubs.	Downloads or views in social web sites	Scholarly		Yes	Manual or from site owner?
All pubs.	Forum citations	Scholarly; educ.; public	Weak associations with	Yes	Commercial altmetric provider

			citations		
All pubs.	Google Scholar citations	Scholarly	Correlations with WoS and Scopus citations	Yes	Publish or Perish software for individual cases; manual
All pubs.	Mendeley bookmarks	Scholarly; educational?	Correlations with citations	Yes	Automatic collection via API
All pubs.	Other bookmarks	Scholarly; educational?			Manual
All pubs.	Patents (Google patent search)	Commercial		No	Manual
All pubs.	Tweets	Mainly scholarly	Weak association with citations	Yes	Buy from altmetrics providers?
All pubs.	Web or URL citations	Educational, scholarly	Correlations with citations	Yes?	Automatic collection via Bing API
All pubs.	Web presentation mentions	Educational, scholarly	Correlations with citations	Yes?	Automatic collection via Bing API
All pubs.	Web syllabus mentions	Educational	Correlations with citations	Yes?	Automatic collection via Bing API
Soft-ware	WoS citations	Scholarly	None		WoS
Videos	Views, comments, sentiment	Scholarly; educ.; cultural	None	Yes	Automatic collection via API in some cases

References

Abbas, A. M. (2011). Weighted indices for evaluating the quality of research with multiple authorship. Scientometrics, 88(1), 107-131.

Abbott, A. (1988). The system of professions : an essay on the division of expert labor. Chicago: University of Chicago Press.

Abelson, P. (1990). Mechanisms for evaluating scientific information and the role of peer review. Journal of the American Society for Information Science, 41(3), 216–222. doi:10.1002/(SICI)1097-4571(199004)41:3<216::AID-ASI13>3.0.CO;2-6

Abramo, G., & D’Angelo, C. A. (2011). Evaluating research: from informed peer review to bibliometrics. Scientometrics, 87, 499–514. doi:10.1007/s11192-011-0352-7

Abramo, G., Cicero, T., & D’Angelo, C. A. (2011). The dangers of performance-based research funding in non-competitive higher education systems. Scientometrics, 87(3), 641–654. doi:10.1007/s11192-011-0355-4

Abramo, G., Cicero, T., & D’Angelo, C. A. (2012a). How important is choice of the scaling factor in standardizing citations? Journal of Informetrics, 6(4), 645-654.

Abramo, G., Cicero, T., & D’Angelo, C. A. (2012b). Revisiting the scaling of citations for research assessment. Journal of Informetrics, 6(4), 470–479. doi:10.1016/j.joi.2012.03.005

Abramo, G., D’Angelo, C. A., & Cicero, T. (2012). What is the appropriate length of the publication period over which to assess research performance? Scientometrics, 93(3), 1005–1017. doi:10.1007/s11192-012-0714-9

Abramo, G., D’Angelo, C. A., & Di Costa, F. (2010). Citations versus journal impact factor as proxy of quality: Could the latter ever be preferable? Scientometrics, 84(3), 821-833.

Abramo, G., D’Angelo, C. A., & Di Costa, F. (2011a). National research assessment exercises: a comparison of peer review and bibliometrics rankings. Scientometrics. doi:10.1007/s11192-011-0459-x

Abramo, G., D'Angelo, C. A., & Di Costa, F. (2011b). National research assessment exercises: the effects of changing the rules of the game during the game. Scientometrics, 88(1), 229–238. doi:10.1007/s11192-011-0373-2

Abramo, G., D'Angelo, C. A., & Rosati, F. (2013). The importance of accounting for the number of co-authors and their order when assessing research performance at the individual level in the life sciences. Journal of Informetrics, 7(1), 198-208.

Abramo, G., D'Angelo, C. A., & Viel, F. (2011). The field-standardized average impact of national research systems compared to world average: the case of Italy. Scientometrics, 88(2), 599–615. doi:10.1007/s11192-011-0406-x

Abrizah, A., & Thelwall, M. (2014), Can the impact of non-Western academic books be measured? An investigation of Google Books and Google Scholar for Malaysia. Journal of the Association for Information Science and Technology. doi: 10.1002/asi.23145

ACUMEN Portfolio: Guidelines for Good Evaluation Practice (2014). The ACUMEN Consortium. http://research-acumen.eu/wp-content/uploads/D6.14-Good-Evaluation-Practices.pdf

Adams, J., Gurney, K., & Jackson, L. (2008). Calibrating the zoom—A test of Zitt's hypothesis. Scientometrics, 75(1), 81-95.

Adie, E., Roe, W. (2013). Altmetric: Enriching scholarly content with article-level discussion and metrics. Learned Publishing, 26: 11–17. http://figshare.com/articles/Enriching_s cholarly_content_with_article_level_discussion_and_metrics/105851.

Aguillo, I. F., Granadino, B., Ortega, J. L., & Prieto, J. A. (2006). Scientific research activity and communication measured with cybermetrics indicators. Journal of the American Society for Information Science and Technology, 57(10), 1296-1302.

Aksnes, D. W. (2003). A macro study of self-citation. Scientometrics, 56(2), 235-246.

Aksnes, D. W., & Rip, A. (2009). Researchers' perceptions of citations. Research Policy, 38(6), 895-905.

Aksnes, D. W., & Sivertsen, G. (2004). The effect of highly cited papers on national citation indicators. Scientometrics, 59(2), 213-224.

Aksnes, D. W., & Taxt, R. E. (2004). Peer reviews and bibliometric indicators: a comparative study at a Norwegian university. Research Evaluation, 13(1), 33–41. doi:10.3152/147154404781776563

Aksnes, D. W., Schneider, J. W., & Gunnarsson, M. (2012). Ranking national research systems by citation indicators. A comparative analysis using whole and fractionalised counting methods. Journal of Informetrics, 6(1), 36-43.

Akst, J. (2010). I Hate Your Paper. The Scientist, 24(8), 36–41.

Albarrán, P., Crespo, J. A., Ortuño, I., & Ruiz-Castillo, J. (2011). The skewness of science in 219 sub-fields and a number of aggregates. Scientometrics, 88(2), 385-397.

Alhoori, H., & Furuta, R. (2014). Do altmetrics follow the crowd or does the crowd follow altmetrics? In: Proceedings of the IEEE/ACM Joint Conference on Digital Libraries (JCDL 2014). Los Alamitos: IEEE Press. http://people.tamu.edu/~alhoori/publications/alhoori2014jcdl.pdf

Alldred, P., & Miller, T. (2007). Measuring what’s valued or valuing what’s measured? Knowledge production and the Research Assessment Exercise. In: Gillies, V., & Lucey, H. (eds.), Power, Knowledge and the Academy: The Institutional Is Political (pp. 147-167). Basingstoke: Palgrave.

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(6), e5910.

Allen, N., & Heath, O. (2013). Reputations and Research Quality in British Political Science: The Importance of Journal and Publisher Rankings in the 2008 RAE. The British Journal of Politics & International Relations, 15(1), 147–162. doi:10.1111/1467-856X.12006

Almind, T. C., & Ingwersen, P. (1997). Informetric analyses on the World Wide Web: Methodological approaches to “Webometrics”. Journal of Documentation, 53(4), 404-426.

Alonso, S., Cabrerizo, F. J., Herrera-Viedma, E., & Herrera, F. (2009). h-Index: A review focused in its variants, computation and standardization for different scientific fields. Journal of Informetrics, 3(4), 273-289.

Amara, N., & Landry, R. (2012). Counting citations in the field of business and management: Why use Google Scholar rather than the Web of Science. Scientometrics, 93(3), 553-581.

Anagnostou, P., Capocasa, M., Milia, N., & Bisol, G. D. (2013). Research data sharing: Lessons from forensic genetics. Forensic Science International: Genetics, 7(6), e117-e119.

Angus, E., Thelwall, M., & Stuart D. (2008). General patterns of tag usage among university groups in Flickr, Online Information Review, 32(1), 89-101.

Archambault, É., & Larivière, V. (2009). History of the journal impact factor: Contingencies and consequences. Scientometrics, 79(3), 635-649.

Archambault, É., Campbell, D., Gingras, Y., & Larivière, V. (2009). Comparing bibliometric statistics obtained from the Web of Science and Scopus. Journal of the American Society for Information Science and Technology, 60(7), 1320-1326.

Archambault, É., Vignola-Gagné, É., Côté, G., Larivière, V., & Gingras, Y. (2006). Benchmarking scientific output in the social sciences and humanities: The limits of existing databases. Scientometrics, 68(3), 329-342.

Armstrong, P. W., Caverson, M. M., Adams, L., Taylor, M., & Olley, P. M. (1997). Evaluation of the Heart and Stroke Foundation of Canada Research Scholarship Program: research productivity and impact. The Canadian Journal of Cardiology, 13(5), 507–16. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9179090

Åström, F., & Hansson, J. (2013). How implementation of bibliometric practice affects the role of academic libraries. Journal of Librarianship and information Science, 45(4), 316-322.

Bai, X. (2011). Predicting consumer sentiments from online text. Decision Support Systems, 50(4), 732-742.

Bakkalbasi, N., Bauer, K., Glover, J., & Wang, L. (2006). Three options for citation tracking: Google Scholar, Scopus and Web of Science. Biomedical Digital Libraries, 3(1), 7.

Bar-Ilan, J. (2008a). Informetrics at the beginning of the 21st century - A review. Journal of Informetrics, 2(1), 1-52.

Bar-Ilan, J. (2008b). Which h-index? A comparison of WoS, Scopus and Google Scholar. Scientometrics, 74(2), 257-271.

Bar-Ilan, J. (2010). Web of Science with the Conference Proceedings Citation Indexes: The case of computer science. Scientometrics, 83(3), 809-824.

Bar-Ilan, J. (2012). JASIST@Mendeley. Presented at the altmetrics12 workshop of the ACM Web Science Conference. http://altmetrics.org/altmetrics12/bar-ilan/.

Bar-Ilan, J. (2014). Astrophysics publications on arXiv, Scopus and Mendeley: A case study. Scientometrics, 100(1), 217-225.

Bar-Ilan, J., Haustein, S., Peters, I., Priem, J., Shema, H., & Terliesner, J. (2012): Beyond citations: Scholars' visibility on the social Web. 17th International Conference on Science and Technology Indicators, Montréal, Canada (pp. 98–109). Retrieved from http://arxiv.org/abs/1205.5611

Bartol, T., Budimir, G., Dekleva-Smrekar, D., Pusnik, M., & Juznic, P. (2014). Assessment of research fields in Scopus and Web of Science in the view of national research evaluation in Slovenia. Scientometrics, 98(2), 1491-1504.

Batista, P. D., Campiteli, M. G., & Kinouchi, O. (2006). Is it possible to compare researchers with different scientific interests? Scientometrics, 68(1), 179-189.

Bauer, K., & Bakkalbasi, N. (2005). An examination of citation counts in a new scholarly communication environment. D-Lib Magazine, 11(9), http://www.dlib.org/dlib/september05/bauer/09bauer.html

Becher, T. (1989). Academic tribes and territories : intellectual enquiry and the cultures of disciplines. Milton Keynes, England; Bristol PA. USA: Society for Research into Higher Education ; Open University Press.

Beel, J., & Gipp, B. (2010a). Academic search engine spam and Google Scholar's resilience against it. Journal of Electronic Publishing, 13(3).

Beel, J., & Gipp, B. (2010b). On the robustness of Google Scholar against spam. Paper presented at the HT'10 - Proceedings of the 21st ACM Conference on Hypertext and Hypermedia, 297-298.

Bence, V., & Oppenheim, C. (2004). The influence of peer review on the research assessment exercise. Journal of Information Science, 30(4), 347-368.

Benda, W. G. G., & Engels, T. C. E. (2010). The predictive validity of peer review : A selective review of the judgmental forecasting qualities of peers, and implications for innovation in science. International Journal of Forecasting, (January), 1–46.

Bensman, S. J. (2007). Garfield and the impact factor. Annual Review of Information Science and Technology, 41(1), 93-155.

Bergstrom, C. T. (2007). Eigenfactor: Measuring the value and prestige of scholarly journals. College and Research Libraries News, 68(5), 314–316.

Birkholz, J., & Wang, S. (2011). Who are we talking about?: The validity of online metrics for commenting on science. Paper presented in: altmetrics11: Tracking scholarly impact on the social Web. An ACM Web Science Conference 2011 Workshop, Koblenz (Germany), 14-15. http://altmetrics.org/workshop2011/birkholz-v0

Bogers, T., & Bosch, A. v. d. (2008). Recommending scientific articles using CiteULike. Proceedings of the 2008 ACM conference on recommender systems (RecSys '08) (pp. 287-290). New York, NY: ACM.

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(1), 136-149.

Bollen, J., Rodriquez, M. A., & Van de Sompel, H. (2006). Journal status. Scientometrics, 69(3), 669-687.

Bollen, J., Van De Sompel, H., & Rodriguez, M. A. (2008). Towards usage-based impact metrics: First results from the MESUR project. Paper presented at the Proceedings of the ACM International Conference on Digital Libraries, Pittsburgh, Pennsylvania, USA, 231-240.

Bollen, J., Van de Sompel, H., Hagberg, A., & Chute, R. (2009). A principal component analysis of 39 scientific impact measures. PLOS ONE, 4(6), e6022.

Bollen, J., Van De Sompel, H., Smith, J. A., & Luce, R. (2005). Toward alternative metrics of journal impact: A comparison of download and citation data. Information Processing & Management, 41(6), 1419-1440.

Boltanski, L., & Thévenot, L. (2006). On justification: Economies of worth. Princeton, NJ: Princeton University Press.

Borgman, C. (2007). Scholarship in the Digital Age. Cambridge: MIT Press.

Borgman, C. L. (2012). The conundrum of sharing research data. Journal of the American Society for Information Science and Technology, 63(6), 1059-1078.

Borgman, C., & Furner, J. (2002). Scholarly communication and bibliometrics. Annual Review of Information Science and Technology, 36, Medford, NJ: Information Today Inc., pp. 3-72.

Bornmann, L. & Daniel, H. (2008). What do citation counts measure? A review of studies on citing behavior. Journal of Documentation, 64(1), 45–80. doi:10.1108/00220410810844150

Bornmann, L. (2011a). Scientific Peer Review. Annual Review of Information Science and Technology, 45(5), 197–245.

Bornmann, L. (2011b). The Hawthorne effect in journal peer review. Scientometrics, 91(3), 857–862. doi:10.1007/s11192-011-0547-y

Bornmann, L., & Daniel, H.-D. (2005). Selection of research fellowship recipients by committee peer review. Reliability, fairness and predictive validity of Board of Trustees' decisions. Scientometrics, 63(2), 297–320.

Bornmann, L., & Daniel, H.-D. (2006). Selecting scientific excellence through committee peer review – A citation analysis of publications previously published to approval or rejection of post-doctoral research fellowship applicants. Scientometrics, 68(3), 427–440.

Bornmann, L., & Daniel, H.-D. (2008a). Selecting manuscripts for a high-impact journal through peer review: A citation analysis of communications that were accepted by Angewandte Chemie International Edition, or rejected but published elsewhere. Journal of the American Society for Information Science and Technology, 59(11), 1841–1852. doi:10.1002/asi.20901

Bornmann, L., & Daniel, H.-D. (2008b). The effectiveness of the peer review process: inter-referee agreement and predictive validity of manuscript refereeing at Angewandte Chemie. Angewandte Chemie (International Ed. in English), 47(38), 7173–8. doi:10.1002/anie.200800513

Bornmann, L., & Daniel, H.-D. (2010a). How Long is the Peer Review Process for Journal Manuscripts?
 A Case Study on <I>Angewandte Chemie International Edition</I>. CHIMIA International Journal for Chemistry, 64, 72–77. doi:10.2533/chimia.2010.72

Bornmann, L., & Daniel, H.-D. (2010b). The manuscript reviewing process: Empirical research on review requests, review sequences, and decision rules in peer review. Library & Information Science Research, 32, 5–12. doi:10.1016/j.lisr.2009.07.010

Bornmann, L., & Daniel, H.-D. (2010c). The Usefulness of Peer Review for Selecting Manuscripts for Publication: A Utility Analysis Taking as an Example a High-Impact Journal. PLoS ONE, 5, e11344. doi:10.1371/journal.pone.0011344

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(2), 286-291.

Bornmann, L., & Mungra, P. (2011). Improving peer review in scholarly journals, 37(May), 41–43. doi:10.4103/0256-4602.60162.3

Bornmann, L., & Mutz, R. (2011). Further steps towards an ideal method of measuring citation performance: The avoidance of citation (ratio) averages in field-normalization. Journal of Informetrics, 5(1), 228-230.

Bornmann, L., De Moya Anegón, F., & Leydesdorff, L. (2012). The new excellence Indicator in the World Report of the SCImago Institutions Rankings 2011. Journal of Informetrics, 6(2), 333-335.

Bornmann, L., Marx, W., Schier, H., Rahm, E., Thor, A., & Daniel, H. (2009). Convergent validity of bibliometric Google Scholar data in the field of chemistry-citation counts for papers that were accepted by angewandte chemie international edition or rejected but published elsewhere, using Google Scholar, Science Citation Index, Scopus, and chemical abstracts. Journal of Informetrics, 3(1), 27-35.

Bornmann, L., Mutz, R., Neuhaus, C., & Daniel, H. D. (2008). Citation counts for research evaluation: Standards of good practice for analyzing bibliometric data and presenting and interpreting results. Ethics in Science and Environmental Politics, 8(1), 93-102.

Bornmann, L., Wallon, G., & Ledin, A. (2008). Does the committee peer review select the best applicants for funding? An investigation of the selection process for two European molecular biology organization programmes. PloS One, 3(10), e3480. doi:10.1371/journal.pone.0003480

Bornstein, R. F. (2011). The predictive validity of peer review: A neglected issue. Behavioral and Brain Sciences, 14(01), 138–139. doi:10.1017/S0140525X00065717

Borrego, A., & Fry, J. (2012). Measuring researchers’ use of scholarly information through social bookmarking data: A case study of BibSonomy. Journal of Information Science, 38(3), 297-308.

Braun, T. (2012). Editorial. Scientometrics, 92(2), 207-208.

Braun, T., Glänzel, W., & Schubert, A. (2006). A Hirsch-type index for journals. Scientometrics, 69(1), 169-173.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. Computer Networks and ISDN Systems, 30(1), 107-117.

Brody, T., Harnad, S., & Carr, L. (2006). Earlier web usage statistics as predictors of later citation impact. Journal of the American Society for Information Science and Technology, 57(8), 1060-1072.

Buela-Casal, G., & Zych, I. (2012). What do the scientists think about the impact factor? Scientometrics, 92(2), 281-292.

Burrows, R. (2012). Living with the h-index? Metric assemblages in the contemporary academy. The Sociological Review, 60(2), 355-372.

Butler, L. (2003). Explaining Australia’s increased share of ISI publications - the effects of a funding formula based on publication counts. Research Policy, 32(1), 143-155.

Butler, L. (2005). What happens when funding is linked to publication counts? In: Moed, H., Glänzel, W., & Schmoch, U. (Eds.), Handbook of quantitative science and technology research (pp. 389-405). Dordrecht: Springer.

Butler, L. (2007). Assessing university research: a plea for a balanced approach. Science and Public Policy, 34(8), 565–574. doi:10.3152/030234207X254404

Butler, L. (2008). ICT assessment: Moving beyond journal outputs. Scientometrics, 74(1), 39-55.

Butler, L., & McAllister, I. (2009). Metrics or Peer Review? Evaluating the 2001 UK Research Assessment Exercise in Political Science. Political Studies Review, 7(1), 3–17. doi:10.1111/j.1478-9299.2008.00167.x

Butler, L., & McAllister, I. (2011). Evaluating University Research Performance Using Metrics. European Political Science, 10(1), 44–58. doi:10.1057/eps.2010.13

Cabanac, G. (2012). Shaping the Landscape of Research in Information Systems From the Perspective of Editorial Boards : A Scientometric Study of 77 Leading Journals. Journal of the American Society for Information Science and Technology, 63(5), 977–996. doi:10.1002/asi

Cabanac, G., & Preuss, T. (2013). Capitalizing on order effects in the bids of peer-reviewed conferences to secure reviews by expert referees. Journal of the American Society for Information Science and Technology, 64(2), 405–415. doi:10.1002/asi.22747

Cabezas-Clavijo, A., Robinson-García, N., Escabias, M., & Jiménez-Contreras, E. (2013). Reviewers' ratings and bibliometric indicators: hand in hand when assessing over research proposals? PloS One, 8(6), e68258. doi:10.1371/journal.pone.0068258

Cabezas-Clavijo, A., Robinson-García, N., Torres-Salinas, D., Jiménez-Contreras, E., Mikulka, T., Gumpenberger, C., Wemisch, A., & Gorraiz, J. (2013). Most borrowed is most cited? Library loan statistics as a proxy for monograph selection in citation indexes. In: Proceedings of 14th International Conference of the International Society for Scientometrics and Informetrics, Vienna, Austria, Vol. 2, pp. 1237-1252. Retrieved from http://arxiv.org/ftp/arxiv/papers/1305/1305.1488.pdf.

Calver, M. C., & Bradley, J. S. (2009). Should we use the mean citations per paper to summarise a journal's impact or to rank journals in the same field? Scientometrics, 81(3), 611-615.

Campanario, J. M. (2011). Empirical study of journal impact factors obtained using the classical two-year citation window versus a five-year citation window. Scientometrics, 87(1), 189-204.

Chakraborty, N. (2012). Activities and reasons for using social networking sites by research scholars in NEHU: A study on Facebook and ResearchGate. Planner-2012, 19-27. http://ir.inflibnet.ac.in/bitstream/1944/1666/1/3.pdf

Chandler, J., Barry, J., & Clark, H. (2002). Stressing academe: The wear and tear of the new public management. Human Relations, 55(9), 1051-1069.

Chavan, V. S., & Ingwersen, P. (2009). Towards a data publishing framework for primary biodiversity data: Challenges and potentials for the biodiversity informatics community. BMC Bioinformatics, 10 (SUPPL.14). S2: Doi: 10.1186/1471-2105-10-S14-S2

Chen, X. (2010). Google Scholar's dramatic coverage improvement five years after debut. Serials Review, 36(4), 221–226.

Chen, X. (2012). Google Books and WorldCat: A comparison of their content. Online Information Review, 36(4), 507-516.

Chew, M., Villanueva, E. V., & Van Der Weyden, M. B. (2007). Life and times of the impact factor: retrospective analysis of trends for seven medical journals (1994-2005) and their Editors' views. Journal of the Royal Society of Medicine, 100(3), 142-150.

Chu, H., & Krichel, T. (2007). Downloads vs. citations in economics: Relationships, contributing factors and beyond. Paper presented at the Proceedings of ISSI 2007 - 11th International Conference of the International Society for Scientometrics and Informetrics, pp. 207–215, Madrid, Spain, June 25–27. http://eprints.rclis.org/11085/1/DownloadsVsCitations.pdf

Chubin, D. E., & Hackett, E. J. (1990). Peerless science : peer review and U.S. science policy. Albany, N.Y.: State University of New York Press. Retrieved from http://books.google.nl/books?id=Xfsh6D29WoIC&pg=PA129&lpg=PA129&dq=peerless+science+hackett&source=bl&ots=gks5p15I6p&sig=RwupXl4tvFtOw8nrHBidGipJqR0&hl=en&sa=X&ei=6gP7Tq_iHMzz-gbEmrW3AQ&ved=0CCoQ6AEwAA#v=onepage&q=peerless science hackett&f=false

Cicchetti, D. V. (1991). The reliability of peer review for manuscript and grant submissions: A cross-disciplinary investigation. Behavioral and Brain Sciences, 14, 119–186.

Clerides, S., Pashardes, P., & Polycarpou, A. (2011). Peer Review vs Metric-based Assessment: Testing for Bias in the RAE Ratings of UK Economics Departments. Economica, 78(311), 565–583. doi:10.1111/j.1468-0335.2009.00837.x

Cole, J. R., & Cole, S. (1981). Peer Review in the National Science Foundation. Phase Two of a Study. Washington D.C.: National Academy Press.

Cole, S., Rubin, L., & Cole, J. R. (1978). Peer Review in the National Science Foundation. Phase One of a Study. Washington D.C.: National Academy Press.

Cole, S., Simon, G., & others. (1981). Chance and consensus in peer review. Science, 214(4523), 881. Retrieved from http://www.sciencemag.org/content/214/4523/881.short

Colliander, C. (in press). A novel approach to citation normalization: A similarity-based method for creating reference sets. Journal of the Association for Information Science and Technology.

Colliander, C., & Ahlgren, P. (2011). The effects and their stability of field normalization baseline on relative performance with respect to citation impact: A case study of 20 natural science departments. Journal of Informetrics, 5(1), 101-113.

Collins, R. (1998). The Sociology of Philosophies: A Global Theory of Intellectual Change. Belknap Press of Harvard University Press. Retrieved from http://www.amazon.com/dp/0674001877

Colwell, R. et al., (2012). Informing Research Choices: Indicators and Judgment. The Expert Panel on Science Performance and Research Funding. Ottawa: Council of Canadian Academics.

Colwell, R., Blouw, M., Butler, L., Cozzens, S. E., Feller, I., Gingras, Y., … & Woodward, R. (2012). Informing Research Choices: Indicators and Judgment. The Expert Panel on Science Performance and Research Funding (p. 142). Ottawa.

Colyvas, J. A. (2012). Performance Metrics as Formal Structures and through the Lens of Social Mechanisms: When Do They Work and How Do They Influence? American Journal of Education, 118(2), 167-197.

Cooper, M. D., & McGregor, G. F. (1994). Using article photocopy data in bibliographic models for journal collection management. Library Quarterly, 64(4), 386-413.

Costas, R., Van Leeuwen, T. N., & Bordons, M. (2010). Self-citations at the meso and individual levels: Effects of different calculation methods. Scientometrics, 82(3), 517-537.

Costas, R., Zahedi, Z., & Wouters, P. (2014). Do altmetrics correlate with citations? Extensive comparison of altmetric indicators with citations from a multidisciplinary perspective. arXiv preprint arXiv:1401.4321.

Cozzens, S., & Melkers, J. (1997). Use and Usefulness of Performance Measurement in State Science and Technology Programs. Policy Studies Journal, 25(3), 425-35.

Craig, R., Amernic, J., & Tourish, D. (2014). Perverse Audit Culture and Accountability of the Modern Public University. Financial Accountability & Management, 30(1), 1-24.

Cronin, B. (2001a). Bibliometrics and beyond: Some thoughts on web-based citation analysis. Journal of Information Science, 27(1), 1-7.

Cronin, B. (2001b). Hyperauthorship: A postmodern perversion or evidence of a structural shift in scholarly communication practices? Journal of the American Society for Information Science and Technology, 52(7), 558-569.

Cronin, B., & Sugimoto, C. R. (Eds.). (2014). Beyond Bibliometrics: Harnessing Multidimensional Indicators of Scholarly Impact. Cambridge, MA: MIT Press.

Cronin, B., & Weaver, S. (1995). The praxis of acknowledgement: from bibliometrics to influmetrics. Revista Española de Documentación Científica, 18(2): 172-177.

Cronin, B., Snyder, H.W., Rosenbaum, H., Martinson, A., & Callahan, E. (1998). Invoked on the web. Journal of the American Society for Information Science, 49(14), 1319–1328

Crotty, D. (2012). Life after publication - post-publication peer review. Biochemist, 34(4), 26-28. http://www.biochemist.org/bio/03404/0026/034040026.pdf

Crowston, K., Annabi, H., Howison, J. (2003). Defining open source software project success. In proceedings of the 24th International Conference on Information Systems (ICIS), Seattle, Washington, USA, pp. 327-340.

Crowston, K., Annabi, H., Howison, J., & Masango, C. (2004). Towards a portfolio of FLOSS project success measures, the 4th workshop on Open Source Software engineering, International Conference on Software Engineering, May 25, Edinburgh, Scotland.

da Silva, J. A. T. (2013). The need for post-publication peer review in plant science publishing. Frontiers in Plant Science, (4 December): Doi: 10.3389/fpls.2013.00485.

Dahler-Larsen, P. (2012). The evaluation society, Stanford CA: Stanford Business Books.

Dahler-Larsen, P. (2014). Constitutive Effects of Performance Indicators. Public Management Review, 16(7), 969-986.

Daniel, H.-D. (1993). Guardians of science : fairness and reliability of peer review. Weinheim;New York: VCH.

Darmoni, S. J., Roussel, F., Benichou, J., Faure, G. C., Thirion, B., & Pinhas, N. (2000). Reading factor as a credible alternative to impact factor: A preliminary study. Technology and Health Care, 8 (3-4), 174–175.

Darnton, R. (2013). The national digital public library is launched. New York Review of Books, 60(7), Retrieved July 02, 2013, from http://www.nybooks.com/articles/archives/2013/apr/25/national-digital-public-library-launched/

Daston, L., & Galison, P. (2007). Objectivity. New York: Zone Books.

Davies, J., & Merchant, G. (2007). Looking from the inside out: Academic blogging as new literacy. In C. Lankshear & M. Knobel (Eds.), A New Literacies Sampler (pp. 167–198). NewYork: Peter Lang. http://www.geocities.ws/cornerbrookresearch07/Sampler.pdf#page=177

Davis, P. M. (2008). Eigenfactor: Does the principle of repeated improvement result in better estimates than raw citation counts? Journal of the American Society for Information Science and Technology, 59(13), 2186-2188.

De Boer, H. F., Enders, J., & Leisyte, L. (2007). Public sector reform in Dutch higher education: The organizational transformation of the university. Public Administration, 85(1), 27-46.

Deem, R., Hillyard, S., & Reed, M. (2007). Knowledge, Higher Education and the New Managerialism: The Changing Management of UK Universities. Oxford: Oxford University Press.

De Groote, S. L., & Raszewski, R. (2012). Coverage of Google scholar, Scopus, and Web of Science: A case study of the h-index in nursing. Nursing Outlook, 60(6), 391-400.

Demšar, F., & Južnič, P. (2013). Transparency of research policy and the role of librarians. Journal of Librarianship and Information Science, 46, 139-147.

De Rijcke, S., & Rushforth, A.D. (in press). To intervene, or not to intervene, is that the question? On the role of scientometrics in research evaluation. Journal of the Association for Information Science and Technology.

De Rijcke, S., Wallenburg, I., Bal, R. Wouters, P. (2014). Comparing Comparisons. On Rankings and Accounting in Hospitals and Universities. In: Guggenheim, M., Deville, J., & Hrdlickova, Z. (eds.) Practicing Comparison: Revitalizing the Comparative Act. Manchester: Mattering Press.

De Solla Price, D. (1981). Multiple authorship. Science, 212, 986.

De Winter, J. C. F., Zadpoor, A. A., & Dodou, D. (2014). The expansion of Google Scholar versus Web of Science: A longitudinal study. Scientometrics, 98(2), 1547-1565.

Delgado López-Cózar, E., Robinson-García, N., & Torres-Salinas, D. (2014). The Google Scholar experiment: How to index false papers and manipulate bibliometric indicators. Journal of the Association for Information Science and Technology, 65(3), 446-454.

Derrick, G. E., & Pavone, V. (2013). Democratising research evaluation: Achieving greater public engagement with bibliometrics-informed peer review. Science and Public Policy, 40(5), 563–575. doi:10.1093/scipol/sct007

Derrick, G. E., Haynes, A., Chapman, S., & Hall, W. D. (2011). The association between four citation metrics and peer rankings of research influence of Australian researchers in six fields of public health. PloS One, 6(4), e18521. doi:10.1371/journal.pone.0018521

Derrick, G., & Gillespie, J. (2013) "A number you just can’t get away from”: Characteristics of Adoption and the Social Construction of Metric Use by Researchers. STI 18th International Conference on Science and Technology Indicators. Berlin.

Desai, T., Shariff, A., Shariff, A., Kats, M., Fang, X., Christiano, C., & Ferris, M. (2012). Tweeting the meeting: An in-depth analysis of Twitter activity at kidney week 2011. PLoS ONE, 7(7)

Donovan, C. (2007). The qualitative future of research evaluation. Science and Public Policy, 34(8), 585–597. doi:10.3152/030234207X256538

Donovan, C., & Butler, L. (2007). Testing novel quantitative indicators of research “quality,” esteem and “user engagement:” An economics pilot study. Research Evaluation, 16(4), 231–242.

Drott, C. M. (1995). Reexamining the role of conference papers in scholarly communication. Journal of the American Society for Information Science, 46(4), 299–305.

Duin, D., King, D., & van den Besselaar, P. (2012). Identifying audiences of e-infrastructures-tools for measuring impact. PloS ONE, 7(12), e50943.

Dullaart, C. (2014). High Retention, Slow Delivery. [Art piece: 2.5 million Instagram followers bought and distributed to artists. See e.g. http://jeudepaume.espacevirtuel.org/ http://dismagazine.com/dystopia/67039/constant-dullaart-100000-followers-for-everyone/]

Duy, J., & Vaughan, L. (2006). Can electronic journal usage data replace citation data as a measure of journal use? An empirical examination. Journal of Academic Librarianship, 32(5), 512-517.

Egghe, L. (2006). Theory and practise of the g-index. Scientometrics, 69(1), 131-152.

Egghe, L. (2008). Mathematical theory of the h- and g-index in case of fractional counting of authorship. Journal of the American Society for Information Science and Technology, 59(10), 1608-1616.

Egghe, L. (2010). The Hirsch index and related impact measures. Annual Review of Information Science and Technology, 44(1), 65-114.

Egghe, L., Rousseau, R., & Van Hooydonk, G. (2000). Methods for accrediting publications to authors or countries: Consequences for evaluation studies. Journal of the American Society for Information Science, 51(2), 145-157.

Eisen, J. A., MacCallum, C. J., & Neylon, C. (2013). Expert Failure: Re-evaluating Research Assessment. PLoS Biology, 11(10), e1001677. Retrieved from http://dx.plos.org/10.1371/journal.pbio.1001677

Elkana, Y., Lederberg, J., Merton, R. K., Thackray, A., & Zuckerman, H. (1978). Toward a metric of science : the advent of science indicators. New York ;Chichester etc.: Wiley.

Elkins, M. R., Maher, C. G., Herbert, R. D., Moseley, A. M., & Sherrington, C. (2010). Correlation between the journal impact factor and three other journal citation indices. Scientometrics, 85(1), 81-93.

Elton, L. (2000). The UK research assessment exercise: unintended consequences. Higher Education Quarterly, 54(3), 274-283.

Enders, J. (2009). Richard Whitley, Jochen Gläser (eds.), The Changing Governance of the Sciences. The Advent of Research Evaluation Systems. Sociology of the Sciences Yearbook: Springer, Dordrecht, 2007, 26 pp. Minerva, 47(4), 465–468. doi:10.1007/s11024-009-9132-4

Engels, T. C., Ossenblok, T. L., & Spruyt, E. H. (2012). Changing publication patterns in the social sciences and humanities, 2000–2009. Scientometrics, 93(2), 373-390.

Engqvist, L., & Frommen, J. G. (2008). The h-index and self-citations. Trends in Ecology and Evolution, 23(5), 250-252.

Engqvist, L., & Frommen, J. G. (2010). New insights into the relationship between the h-index and self-citations? Journal of the American Society for Information Science and Technology, 61(7), 1514-1516.

Espeland, W. N., & Sauder, M. (2007). Rankings and Reactivity: How Public Measures Recreate Social Worlds. American journal of sociology, 113(1), 1-40.

Espeland, W. N., & Stevens, M. L. (1998). Commensuration as a social process. Annual review of sociology, 24, 313-343.

Etzkowitz, H., & Leydesdorff, L. (2000). The dynamics of innovation: from national systems and “Mode 2” to a triple helix of university-industry-government relations. Research Policy, 29, 109–123.

Ewins, R. (2005). Who are you? Weblogs and academic identity. E-Learning, 2, 368–377.

Eyre-Walker, A., & Stoletzki, N. (2013). The assessment of science: The relative merits of post-publication review, the impact factor, and the number of citations. PLoS Biology, 11(10): e1001675. doi:10.1371/journal.pbio.1001675

Eysenbach, G. (2011). Can tweets predict citations? Metrics of social impact based on Twitter and correlation with traditional metrics of scientific impact. Journal of Medical Internet Research, 13(4). doi:10.2196/jmir.2012

Falagas, M. E., Pitsouni, E. I., Malietzis, G. A., & Pappas, G. (2008). Comparison of PubMed, Scopus, Web of Science, and Google Scholar: Strengths and weaknesses. FASEB Journal, 22(2), 338-342.

Fealing, K., Lane, J., Marburger III, J., & Shipp, S. (eds.) (2011). The Science of Science Policy. A Handbook. Standford, CA: Stanford University Press.

Feller, I. (2009). Performance measurement and the governance of American academic science. Minerva, 47(3), 323-344.

Fenner, M. (2013). What can article-level metrics do for you? PLoS Biology, 11(10): e1001687. doi:10.1371/journal.pbio.1001687

Fowler, J. H., & Aksnes, D. W. (2007). Does self-citation pay? Scientometrics,72(3), 427-437.

Fragkiadaki, E., & Evangelidis, G. (2014). Review of the indirect citations paradigm: Theory and practice of the assessment of papers, authors and journals. Scientometrics, 99(2), 261-288.

Franceschet, M. (2010a). A comparison of bibliometric indicators for computer science scholars and journals on Web of Science and Google Scholar. Scientometrics, 83(1), 243-258.

Franceschet, M. (2010b). The difference between popularity and prestige in the sciences and in the social sciences: A bibliometric analysis. Journal of Informetrics, 4(1), 55-63.

Franceschet, M. (2010c). Journal influence factors. Journal of Informetrics, 4(3), 239-248.

Franceschet, M. (2010d). Ten good reasons to use the eigenfactor metrics. Information Processing and Management, 46(5), 555-558.

Franceschet, M., & Costantini, A. (2011). The first Italian research assessment exercise: A bibliometric perspective. Journal of Informetrics, 5(2), 275–291. doi:10.1016/j.joi.2010.12.002

Franceschini, F., Galetto, M., Maisano, D., & Mastrogiacomo, L. (2012). The success-index: An alternative approach to the h-index for evaluating an individual’s research output. Scientometrics, 92(3), 621-641.

Frandsen, T. F., & Nicolaisen, J. (2010). What is in a name? Credit assignment practices in different disciplines. Journal of Informetrics, 4(4), 608-617.

Frandsen, T. F., & Rousseau, R. (2005). Article impact calculated over arbitrary periods. Journal of the American Society for Information Science and Technology, 56(1), 58-62.

Frodeman, R., & Briggle, A. (2012). The Dedisciplining of Peer Review. Minerva, 50(1), 3–19. doi:10.1007/s11024-012-9192-8

Galam, S. (2011). Tailor based allocations for multiple authorship: A fractional gh-index. Scientometrics, 89(1), 365-379.

Gambardella, A., Nappi, C. A., Peracchi, F., & Bertocchi, G. (2013). Bibliometric Evaluation vs . Informed Peer Review : Evidence from Italy (No. 7739). Bonn.

Garand, J.C., & Giles, M.W. (2011). Ranking scholarly publishers in political science: An alternative approach. PS: Political Science and Politics, 44(2), 375-383.

García-Pérez, M. A. (2010). Accuracy and completeness of publication and citation records in the Web of Science, PsycINFO, and Google Scholar: A case study for the computation of h indices in psychology. Journal of the American Society for Information Science and Technology, 61(10), 2070-2085.

García-Pérez, M. A. (2011). Strange attractors in the Web of Science database. Journal of Informetrics, 5(1), 214-218.

Garfield, E. (1955). Citation Indexes for Science 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. (1996a). Citation indexes for retrieval and research evaluation. Consensus Conference on the Theory and Practice of Research Assessment, Capri.

Garfield, E. (1996b). How can impact factors be improved? British Medical Journal, 313, 411.

Garfield, E. (2006). The history and meaning of the journal impact factor. JAMA, 295(1), 90-93.

Gauffriau, M., & Larsen, P. O. (2005). Counting methods are decisive for rankings based on publication and citation studies. Scientometrics, 64(1), 85-93.

Gauffriau, M., Larsen, P. O., Maye, I., Roulin-Perriard, A., & Von Ins, M. (2007). Publication, cooperation and productivity measures in scientific research. Scientometrics, 73(2), 175-214.

Gauffriau, M., Larsen, P. O., Maye, I., Roulin-Perriard, A., & Von Ins, M. (2008). Comparisons of results of publication counting using different methods. Scientometrics, 77(1), 147-176.

Gavel, Y., & Iselid, L. (2008). Web of Science and Scopus: A journal title overlap study. Online Information Review, 32(1), 8-21.

Gazni, A., Sugimoto, C. R., & Didegah, F. (2012). Mapping world scientific collaboration: Authors, institutions, and countries. Journal of the American Society for Information Science and Technology, 63(2), 323-335.

Geisler, E. (2000). The Metrics of Science and Technology (p. 400). Quorum Books. Retrieved from http://www.amazon.com/Metrics-Science-Technology-Eliezer-Geisler/dp/1567202136

Georghiou, L., et al. (2000). Impact of the Research Assessment Exercise and the Future of Quality Assurance in the Light of Changes in the Research Landscape. Manchester: Policy Research in Engineering, Science and Technology Retrieved November 29, 2014, from https://research.mbs.ac.uk/INNOVATION/Portals/0/docs/raec.pdf

Gianoli, E., & Molina-Montenegro, M. A. (2009). Insights into the relationship between the h-index and self-citations. Journal of the American Society for Information Science and Technology, 60(6), 1283-1285.

Gill, R. (2009). Breaking the silence: The hidden injuries of neo-liberal academia. Secrecy and silence in the research process: Feminist reflections, 228-244.

Giménez-Toledo, E., Tejada-Artigas, C., & Mañana-Rodríguez, J. (2013). Evaluation of scientific books' publishers in social sciences and humanities: Results of a survey. Research Evaluation, 22(1), 64-77.

Glänzel, W. (2011). The application of characteristic scores and scales to the evaluation and ranking of scientific journals. Journal of Information Science,37(1), 40-48.

Glänzel, W. (2013). High-end performance or outlier? Evaluating the tail of scientometric distributions. Scientometrics, 97(1), 13-23.

Glänzel, W., & Moed, H. F. (2002). Journal impact measures in bibliometric research. Scientometrics, 53(2), 171-193.

Glänzel, W., & Schoepflin, U. (1995). A bibliometric study on ageing and reception processes of scientific literature. Journal of information Science, 21(1), 37–53.

Glänzel, W., & Schubert, A. (1988). Characteristic scores and scales in assessing citation impact. Journal of Information Science, 14(2), 123-127.

Glänzel, W., & Schubert, A. (2003). A new classification scheme of science fields and subfields designed for scientometric evaluation purposes. Scientometrics, 56(3), 357-367.

Glänzel, W., & Thijs, B. (2004). The influence of author self-citations on bibliometric macro indicators. Scientometrics, 59(3), 281-310.

Glänzel, W., Debackere, K., Thijs, B., & Schubert, A. (2006a). A concise review on the role of author self-citations in information science, bibliometrics and science policy. Scientometrics, 67(2), 263-277.

Glänzel, W., Schlemmer, B., & Thijs, B. (2003). Better late than never? On the chance to become highly cited only beyond the standard bibliometric time horizon. Scientometrics, 58(3), 571-586.

Glänzel, W., Schlemmer, B., Schubert, A., & Thijs, B. (2006b). Proceedings literature as additional data source for bibliometric analysis. Scientometrics, 68(3), 457-473.

Glänzel, W., Schubert, A., & Czerwon, H. J. (1999). An item-by-item subject classification of papers published in multidisciplinary and general journals using reference analysis. Scientometrics, 44(3), 427-439.

Glänzel, W., Schubert, A., Thijs, B., & Debackere, K. (2011). A priori vs. a posteriori normalisation of citation indicators. The case of journal ranking. Scientometrics, 87(2), 415-424.

Glänzel, W., Thijs, B., & Schlemmer, B. (2004). A bibliometric approach to the role of author self-citations in scientific communication. Scientometrics, 59(1), 63-77.

Glänzel, W., Thijs, B., Schubert, A., & Debackere, K. (2009). Subfield-specific normalized relative indicators and a new generation of relational charts: Methodological foundations illustrated on the assessment of institutional research performance. Scientometrics, 78(1), 165-188.

Glasbey, C., & Horgan, G. (1995). Image analysis for the biological sciences, John Wiley & Sons, Chichester.

Gläser, J. (2010). Concluding Reflections: From Governance to Authority Relations? In: Whitley, R., Gläser, J., & Engwall, L. (eds.), Reconfiguring knowledge production: changing authority relationships in the sciences and their consequences for intellectual innovation (pp. 357-369). Oxford: Oxford University Press.

Gläser, J. (2013). How does Governance change research content? On the possibility of a sociological middle-range theory linking science policy studies to the sociology of scientific knowledge. Berlin: The Technical University Technology Studies Working Papers.

Gläser, J., & Laudel, G. (2001). Integrating scientometric indicators into sociological studies : methodical and methodological problems. Scientometrics, 52(3), 411–434.

Gläser, J., Lange, S., Laudel, G., & Schimank, U. (2010). The Limits of Universality: How Field-specific Epistemic Conditions Affect Authority Relations and their Consequences. In: Whitley, R., Gläser, J., & L. Engwall (Eds.), Reconfiguring knowledge production: changing authority relationships in the sciences and their consequences for intellectual innovation (pp. 291–324). Oxford: Oxford University Press.

Gläser, J., Laudel, G., Hinze, S. & Butler, L. (2002). Impact of evaluation-based funding on the production of scientific knowledge: What to worry about, and how to find out. Report for the German Ministry for Education and Research. Retrieved November 29, 2014, from http://www.laudel.info/wp-content/uploads/2013/pdf/research%20papers/02expertiseglaelauhinbut.pdf

Godlee, F., & Jefferson, T. (1999). Peer Review in Health Sciences. London: BMJ Books.

Gómez-Sancho, J. M., & Mancebón-Torrubia, M. J. (2009). The evaluation of scientific production: Towards a neutral impact factor. Scientometrics, 81(2), 435-458.

González-Albo, B., & Bordons, M. (2011). Articles vs. proceedings papers: Do they differ in research relevance and impact? A case study in the library and information science field. Journal of Informetrics, 5(3), 369-381.

González-Pereira, B., Guerrero-Bote, V. P., & Moya-Anegón, F. (2010). A new approach to the metric of journals' scientific prestige: The SJR indicator. Journal of informetrics, 4(3), 379-391.

Gorraiz, J., & Gumpenberger, C. (2010). Going beyond citations: SERUM - A new tool provided by a network of libraries. LIBER Quarterly, 20(1), 80-93.

Gorraiz, J., Gumpenberger, C., & Purnell, P. J. (2014). The power of book reviews: A simple and transparent enhancement approach for book citation indexes. Scientometrics, 98(2), 841-852.

Gorraiz, J., Gumpenberger, C., & Schlögl, C. (2013). Differences and similarities in usage versus citation behaviours observed for five subject areas. In Paper Presented at the Proceedings of ISSI 2013—14th International Society of Scientometrics and Informetrics Conference (pp. 519-535). Vienna, Austria: AIT Austrian Institute of Technology GmbH Vienna.

Gorraiz, J., Purnell, P. J., & Glänzel, W. (2013). Opportunities for and limitations of the book citation index. Journal of the American Society for Information Science and Technology, 64(7), 1388-1398.

Gregg, M (2006) Feeling Ordinary: Blogging as Conversational Scholarship. Continuum: Journal of Media and Cultural Studies, 20(2),147-160.
http://espace.library.uq.edu.au/eserv.php?pid=UQ:7740&dsID=FeelingOrdinary.htm

Groth, P., & Gurney, T. (2010). Studying scientific discourse on the web using bibliometrics: A chemistry blogging case study. In Proceedings of the WebSci10: Extending the Frontiers of Society On-Line, April 26–27th 2010. Raleigh, NC, USA.
http://journal.webscience.org/308/2/websci10_submission_48.pdf

Guerrero-Bote, V. P., & Moya-Anegón, F. (2012). A further step forward in measuring journals' scientific prestige: The SJR2 indicator. Journal of Informetrics, 6(4), 674-688.

Gurung, R. A. A., Landrum, R. E., & Daniel, D. B. (2012). Textbook use and learning: A North American perspective. Psychology Learning and Teaching, 11(1), 87-98.

Gurung, R. A. R., & Martin, R. C. (2011). Predicting textbook reading: The textbook assessment and usage scale. Teaching of Psychology, 38(1), 22-28.

Haddow, G., & Genoni, P. (2010). Citation analysis and peer ranking of Australian social science journals. Scientometrics, 85(2), 471-487.

Hagen, N. T. (2008). Harmonic allocation of authorship credit: Source-level correction of bibliometric bias assures accurate publication and citation analysis. PLOS ONE, 3(12), e4021.

Hagen, N. T. (2010). Harmonic publication and citation counting: sharing authorship credit equitably– not equally, geometrically or arithmetically. Scientometrics, 84(3), 785-793.

Hagen, N. T. (2013). Harmonic coauthor credit: A parsimonious quantification of the byline hierarchy. Journal of Informetrics, 7(4), 784-791.

Hagen, N. T. (2014a). Counting and comparing publication output with and without equalizing and inflationary bias. Journal of Informetrics, 8(2), 310-317.

Hagen, N. T. (2014b). Reversing the byline hierarchy: The effect of equalizing bias on the accreditation of primary, secondary and senior authors. Journal of Informetrics, 8(3), 618-627.

Hammarfelt, B. (2014). Using altmetrics for assessing research impact in the humanities. Scientometrics. 101(2), 1419-1430.

Hammarfelt, B., & de Rijcke, S. (Forthcoming) Accountability in context: Effects of research evaluation systems on publication practices, disciplinary norms and individual working routines in the faculty of Arts at Uppsala University. Accepted for publication in Research Evaluation.

Haran, B., & Poliakoff, M. (2012). The periodic table of videos. Science, 332, 1046-1047.

Hare, P. G. (2003). The United Kingdom's Research Assessment Exercise: impact on institutions, departments, individuals. Higher Education Management and Policy, 15, 43-62.

Hargens, L. L., & Schuman, H. (1990). Citation counts and social comparisons: Scientists' use and evaluation of citation index data. Social Science Research, 19(3), 205–221.

Harley, D., Acord, S. K., Earl-Novell, S., Lawrence, S., & King, C. J. (2010). Assessing the future landscape of scholarly communication: An exploration of faculty values and needs in seven disciplines. Center for Studies in Higher Education. Retrieved November 29, 2014, from https://escholarship.org/uc/item/15x7385g

Harley, S. (2002). The impact of research selectivity on academic work and identity in UK universities. Studies in Higher Education, 27(2), 187-205.

Harter, S., & Ford, C. (2000). Web-based analysis of E-journal impact: Approaches, problems, and issues. Journal of the American Society for Information Science, 51(13), 1159-76.

Hartley, J. (2006). Reading and writing book reviews across the disciplines. Journal of the American Society for Information Science and Technology, 57(9), 1194-1207.

Harzing, A. & Van der Wal, R. (2008). Google Scholar as a new source for citation analysis. Ethics in Science and Environmental Politics, 8(1), 61-73.

Harzing, A. W. (2010). The publish or perish book. Tarma Software Research.

Harzing, A. W. (2013a). Document categories in the ISI Web of Knowledge: Misunderstanding the social sciences? Scientometrics, 94(1), 23-34.

Harzing, A. W. (2013b). A preliminary test of Google Scholar as a source for citation data: A longitudinal study of Nobel prize winners. Scientometrics, 94(3), 1057-1075.

Harzing, A. W. (2014). A longitudinal study of Google Scholar coverage between 2012 and 2013. Scientometrics, 98(1), 565-575.

Harzing, A. W. & Van der Wal, R. (2009). A Google Scholar h-index for journals: An alternative metric to measure journal impact in economics and business. Journal of the American Society for Information Science and Technology, 60(1), 41-46.

Haustein, S. (2012). Multidimensional journal evaluation. De Gruyter.

Haustein, S. & Siebenlist, T. (2011). Applying social bookmarking data to evaluate journal usage. Journal of Informetrics, 5(3), 446-457.

Haustein, S., Peters, I., Sugimoto, C. R., Thelwall, M. & Larivière, V. (2014). Tweeting biomedicine: An analysis of tweets and citations in the biomedical literature. Journal of the Association for Information Science and Technology, 65(4), 656-669.

Hawkins, C. M., Duszak, R., & Rawson, J. V. (2014). Social media in radiology: Early trends in Twitter microblogging at radiology's largest international meeting. JACR Journal of the American College of Radiology, 11(4), 387-390.

Hemlin, S. (2006). Creative knowledge environments for research groups in biotechnology. The influence of leadership and organizational support in universities and business companies. Scientometrics, 67(1), 121-142.

Hemlin, S., & Rasmussen, S. B. (2006). The Shift in Academic Quality Control. Science, Technology & Human Values, 31(2), 173 –198. doi:10.1177/0162243905283639

Henning, V., & Reichelt, J. (2008). Mendeley - A last.fm for research? IEEE fourth international conference on eScience (eScience '08) (pp. 327-328). Los Alamitos: IEEE.

Henzinger, M., Suñol, J., & Weber, I. (2009). The stability of the h-index. Scientometrics, 84, 465–479.

Henzinger, M., Suñol, J., & Weber, I. (2010). The stability of the h-index. Scientometrics, 84(2), 465-479.

Herrera, G. (2011). Google Scholar users and user behaviors: An exploratory study. College and Research Libraries, 72(4), 316-331.

Hicks, D. (1999). The difficulty of achieving full coverage of international social science literature and the bibliometric consequences. Scientometrics, 44(2), 193-215.

Hicks, D. (2012). Performance-based university research funding systems. Research Policy, 41(2), 251–261.

Hicks, D., & Katz, J. S. (2011). Equity and Excellence in Research Funding. Minerva, 49(2), 137–151. doi:10.1007/s11024-011-9170-6

Hightower, C., & Caldwell, C. (2010). Shifting sands: Science researchers on Google Scholar, Web of Science, and PubMed, with implications for library collections budgets. Issues in Science and Technology Librarianship, 63. DOI: 10.5062/F4V40S4J. http://www.istl.org/10-fall/refereed3.html

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(46), 16569-16572.

Hirsch, J. E. (2010). An index to quantify an individual’s scientific research output that takes into account the effect of multiple coauthorship. Scientometrics, 85(3), 741-754.

Hodge, S. E., & Greenberg, D. A. (1981). Publication credit. Science, 213, 950.

Hoecht, A. (2006). Quality assurance in UK higher education: Issues of trust, control, professional autonomy and accountability. Higher Education, 51(4), 541-563.

Holbrook, B. (2010). Peer review. In R. Frodeman (Ed.), The Oxford handbook of interdisciplinarity (pp. 321–332). Oxford ;;New York : Oxford University Press,.

Holmberg, K., & Thelwall, M. (2014). Disciplinary differences in Twitter scholarly communication, Scientometrics, 101(2), 1027-1042.

Homa, N., Hackathorn, J., Brown, C. M., Garczynski, A., Solomon, E. D., Tennial, R., Sanborn, U. A. &, Gurung, R. A. R. (2013). An analysis of learning objectives and content coverage in introductory psychology syllabi. Teaching of Psychology, 40(3), 169-174.

Hornbostel, S., Böhmer, S., Klingsporn, B., Neufeld, J., & von Ins, M. (2009). Funding of young scientist and scientific excellence. Scientometrics, 79(1), 171–190. doi:10.1007/s11192-009-0411-5

Hu, X. (2009). Loads of special authorship functions: Linear growth in the percentage of "equal first authors" and corresponding authors. Journal of the American Society for Information Science and Technology, 60(11), 2378-2381.

Hu, X., Rousseau, R., & Chen, J. (2010). In those fields where multiple authorship is the rule, the h-index should be supplemented by role-based h-indices. Journal of Information Science, 36(1), 73-85.

Huang, M. H., & Chang, Y. W. (2008). Characteristics of research output in social sciences and humanities: From a research evaluation perspective. Journal of the American Society for Information Science and Technology, 59(11), 1819-1828.

Huang, M. H., & Lin, W. Y. C. (2011). Probing the effect of author self-citations on h index: A case study of environmental engineering. Journal of Information Science, 37(5), 453-461.

Huang, M. H., Lin, C. S., & Chen, D. Z. (2011). Counting methods, country rank changes, and counting inflation in the assessment of national research productivity and impact. Journal of the American Society for Information Science and Technology, 62(12), 2427-2436.

Huang, X., Hawkins, B. A., Lei, F., Miller, G. L., Favret, C., Zhang, R., & Qiao, G. (2012). Willing or unwilling to share primary biodiversity data: Results and implications of an international survey. Conservation Letters, 5(5), 399-406.

Hug, S. E., Ochsner, M., & Daniel, H.-D. (2013). Criteria for assessing research quality in the humanities: a Delphi study among scholars of English literature, German literature and art history. Research Evaluation, 22(5), 369–383. doi:10.1093/reseval/rvt008

Hug, S. E., Ochsner, M., & Daniel, H.-D. (n.d.). A Framework to Explore and Develop Criteria for Assessing Research Quality in the Humanities. International Journal for Education Law and Policy.

Hull, D., Pettifer, S. R., & Kell, D. B. (2008). Defrosting the digital library: Bibliographic tools for the next generation web. PLoS Computational Biology, 4(10), e1000204.

Hunter, J. (2012). Post-publication peer review: Opening up scientific conversation. Frontiers in Computational Neuroscience, (30 August), doi:10.3389/fncom.2012. 00063. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3431010/

Huth, E. J. (1986). Irresponsible authorship and wasteful publication. Annals of Internal Medicine, 104(2), 257-259.

Iglesias, J. E., & Pecharromán, C. (2007). Scaling the h-index for different scientific ISI fields. Scientometrics, 73(3), 303-320.

Ingwersen, P. (1998). The calculation of Web impact factors. Journal of Documentation, 54(2), 236-243.

Ingwersen, P. (2012). The pragmatics of a diachronic journal impact factor. Scientometrics, 92(2), 319-324.

Ingwersen, P., & Chavan, V. (2011). Indicators for the data usage index (DUI): An incentive for publishing primary biodiversity data through global information infrastructure. BMC Bioinformatics, 12(Suppl 15): S3.

Ingwersen, P., & Larsen, B. (2014). Influence of a performance indicator on Danish research production and citation impact 2000–12. Scientometrics, 101(12), 1325-1344.

Ingwersen, P., Larsen, B., Rousseau, R., & Russell, J. (2001). The publication-citation matrix and its derived quantities. Chinese Science Bulletin, 46(6), 524-528.

Jacso, P. (2006). Deflated, inflated and phantom citation counts. Online Information Review, 30(3), 297-309.

Jacsó, P. (2008a). Google Scholar revisited. Online Information Review, 32(1), 102-114.

Jacsó, P. (2008b). The pros and cons of computing the h-index using Scopus. Online Information Review, 32(4), 524-535.

Jacso, P. (2008c). Testing the calculation of a realistic h-index in Google Scholar, Scopus, and Web of Science for F. W. Lancaster. Library Trends, 56(4), 784-815.

Jacsó, P. (2010). Metadata mega mess in Google Scholar. Online Information Review, 34(1), 175-191.

Jacsó, P. (2011). Google Scholar duped and deduped – the aura of "robometrics". Online Information Review, 35(1), 154-160.

James, R. (2010). An assessment of the legibility of Google Books. Journal of Access Services, 7(4), 223-228.

Jansen, B. J., Zhang, M., Sobel, K., & Chowdury, A. (2009). Twitter power: Tweets as electronic word of mouth. Journal of the American Society for Information Science and Technology, 60(11), 2169-2188.

Jayasinghe, U. W., Marsh, H. W., & Bond, N. (2006). A new reader trial approach to peer review in funding research grants : An Australian experiment. Scientometrics, 69(3), 591–606.

Jefferson, T., Wager, E., & Frank Davidoff. (2002). Measuring quality of editorial peer review. Jama-Journal of the American Medical Association, 287(21), 2786–2790.

Jensen, C. B. (2011). Making Lists, Enlisting Scientists. Science Studies, 24(2), 64-84.

Jian, D., & Xiaoli, T. (2013). Perceptions of author order versus contribution among researchers with different professional ranks and the potential of harmonic counts for encouraging ethical co-authorship practices. Scientometrics, 96(1), 277-295.

Jiménez-Contreras, E., de Moya Anegón, F., & López-Cózar, E. D. (2003). The evolution of research activity in Spain: The impact of the National Commission for the Evaluation of Research Activity (CNEAI). Research policy, 32(1), 123-142.

Jobmann, A., Hoffmann, C. P., Künne, S., Peters, I., Schmitz, J., & Wollnik-Korn, G. (2014). Altmetrics for large, multidisciplinary research groups: Comparison of current tools. Bibliometrie-Praxis und Forschung, 3. http://www.bibliometrie-pf.de/article/viewFile/205/258

Jonkers, K., & Derrick, G. E. G. E. (2012). The bibliometric bandwagon: Characteristics of bibliometric articles outside the field literature. Journal of the American Society for Information Science and Technology, 63(4), 829–836. doi:10.1002/asi.22620

Kadriu, A. (2013). Discovering value in academic social networks: A case study in ResearchGate. Paper presented at the Proceedings of the International Conference on Information Technology Interfaces, ITI, 57-62.

Kassirer, J. P. (1994). Peer Review. JAMA, 272(2), 96. doi:10.1001/jama.1994.03520020022005

Kaur, J., Radicchi, F., & Menczer, F. (2013). Universality of scholarly impact metrics. Journal of Informetrics, 7(4), 924-932.

Keenoy, T. (2005). Facing Inwards and Outwards at Once: The liminal temporalities of academic perfomativity. Time & Society, 14(2-3), 303-321.

Keevers, L., Treleaven, L., Sykes, C., & Darcy, M. (2012). Made to measure: Taming practices with results-based accountability. Organization Studies, 33(1), 97-120.

Kelly, A., & Burrows, R. (2012). Measuring the value of sociology ? Some notes on performative metricization in the contemporary academy.

Khabsa, M., & Giles, C. L. (2014). The number of scholarly documents on the public web. PLoS ONE, 9(5): e93949. doi:10.1371/journal.pone.0093949.

Kirkup, G. (2010). Academic blogging: Academic practice and academic identity. London Review of Education, 8(1), 75-84.

Kjellberg, S. (2010). I am a blogging researcher: Motivations for blogging in a scholarly context. First Monday, 15(8).

Kling, R., & McKim, G. (1999). Scholarly communication and the continuum of electronic publishing. Journal of the American Society for Information Science, 50(10), 890-906.

Kling, R., & McKim, G. (2000). Not just a matter of time: Field differences and the shaping of electronic media in supporting scientific communication. Journal of the American Society for Information Science and Technology, 51(14), 1306-1320.

KNAW. (2010). Quality assessment in the design and engineering disciplines. Retrieved from http://www.knaw.nl/Pages/DEF/27/160.bGFuZz1FTkc.html

KNAW. (2011). Quality indicators for research in the humanities. Humanities.

Knorr-Cetina, K., & Mulkay, M. (1983). Science observed: perspectives on the social study of science. London: Sage.

Konkiel, S. (2013). Tracking citations and altmetrics for research data: Challenges and opportunities. Bulletin of the American Society for Information Science and Technology, 39(6), 27-32.

Kosmulski, M. (2011). Successful papers: A new idea in evaluation of scientific output. Journal of Informetrics, 5(3), 481-485.

Kosmulski, M. (2012). The order in the lists of authors in multi-author papers revisited. Journal of Informetrics, 6(4), 639-644.

Kostoff, R. N. (2002). Citation analysis of research performer quality. Scientometrics, 53(1), 49-71.

Kostoff, R. N., & Martinez, W. L. (2005). Is citation normalization realistic?.Journal of information science, 31(1), 57-61.

Kousha, K., & Thelwall, M. (2006). Motivations for URL citations to open access library and information science articles. Scientometrics, 68(3), 501-517.

Kousha, K., & Thelwall, M. (2007a). Google Scholar citations and Google Web/URL citations: A multi-discipline exploratory analysis. Journal of the American Society for Information Science and Technology, 58(7), 1055-1065.

Kousha, K., & Thelwall, M. (2007b). How is science cited on the web? A classification of Google unique web citations. Journal of the American Society for Information Science and Technology, 58(11), 1631-1644.

Kousha, K., & Thelwall, M. (2008a). Assessing the impact of disciplinary research on teaching: An automatic analysis of online syllabuses. Journal of the American Society for Information Science and Technology, 59(13), 2060-2069.

Kousha, K., & Thelwall, M. (2008b). Sources of Google Scholar citations outside the Science Citation Index: A comparison between four science disciplines. Scientometrics, 74(2), 273-294.

Kousha, K., & Thelwall, M. (2009). Google Book Search: Citation analysis for social science and the humanities. Journal of the American Society for Information Science and Technology, 60(8), 1537-1549.

Kousha, K., & Thelwall, M. (2014a). An automatic method for extracting citations from Google Books. Journal of the Association for Information Science and Technology. doi: 10.1002/asi.23170

Kousha, K., & Thelwall, M. (2014b). Web impact metrics for research assessment. In: B. Cronin & C.R. Sugimoto, (Eds), Beyond Bibliometrics: Harnessing Multidimensional Indicators of Scholarly Impact, MIT Press.

Kousha, K., & Thelwall, M. (in press). Can Amazon.com reviews help to assess the wider impacts of books? Journal of the Association for Information Science and Technology.

Kousha, K., Thelwall, M., & Abdoli, M. (2012). The role of online videos in research communication: A content analysis of YouTube videos cited in academic publications. Journal of the American Society for Information Science and Technology, 63(9), 1710-1727.

Kousha, K., Thelwall, M., & Rezaie, S. (2010a). Can the impact of scholarly images be assessed online? An exploratory study using image identification technology. Journal of the American Society for Information Science and Technology, 61(9), 1734-1744.

Kousha, K., Thelwall, M., & Rezaie, S. (2010b). Using the web for research evaluation: The integrated online impact indicator. Journal of Informetrics, 4(1), 124-135.

Kousha, K., Thelwall, M., & Rezaie, S. (2011). Assessing the citation impact of books: The role of Google Books, Google Scholar, and Scopus. Journal of the American Society for Information Science and Technology, 62(11), 2147-2164.

Krücken, G., Blümel, A., & Kloke, K. (2013). The managerial turn in higher education? On the interplay of organizational and occupational change in German academia. Minerva, 51(4), 417-442.

Kryl, D., Allen, L., Dolby, K., Sherbon, B., & Viney, I. (2012). Tracking the impact of research on policy and practice: investigating the feasibility of using citations in clinical guidelines for research evaluation. BMJ Open, 2(2), e000897. Doi:10.1136/bmjopen-2012-000897

Kulkarni, A. V., Aziz, B., Shams, I., & Busse, J. W. (2009). Comparisons of citations in Web of Science, Scopus, and Google Scholar for articles published in general medical journals. JAMA, 302(10), 1092-1096.

Kurtz, M. J., & Bollen, J. (2010). Usage bibliometrics. Annual Review of Information Science and Technology, 44, pp. 1-64.

Kurtz, M. J., Eichhorn, G., Accomazzi, A., Grant, C. S., Murray, S. S., & Watson, J. M. (2000). The NASA astrophysics data system: Overview. Astronomy and Astrophysics Supplement Series, 143(1), 41-59.

Kurtz, M. J., Eichhorn, G., Accomazzi, A., Grant, C., Demleitner, M., Murray, S.S., Martimbeau, N., & Elwell, B. (2005). The bibliometric properties of article readership information. Journal of the American Society for Information Science and Technology, 56(2), 111-128.

Labbé, C. (2010). Ike Antkare, one of the great stars in the scientific firmament. ISSI Newsletter, 6(2), 48-52.

Labbé, C., & Labbé, D. (2013). Duplicate and fake publications in the scientific literature: How many SCIgen papers in computer science? Scientometrics, 94(1), 379-396.

Lamont, M. (2009). How Professors Think: Inside the Curious World of Academic Judgment. Cambridge, MA: Harvard University Press.

Lamont, M. (2012). Toward a Comparative Sociology of Valuation and Evaluation. Annual Review of Sociology, 38, 201-221.

Lange, L. L. (2001). Citation counts of multi-authored papers—First-named authors and further authors. Scientometrics, 52(3), 457-470.

Langfeldt, L. (2004). Expert panels evaluating research: decision-making and sources of bias. Research Evaluation, 13(1), 51–62. doi:10.3152/147154404781776536

Langfeldt, L. (2006). Risk avoidance. The policy challenges of peer review: managing bias, conflict of interests and interdisciplinary assessments. Research Evaluation, 15(1), 31–41. doi:Article

Larivière, V., & Gingras, Y. (2010). Brief communication: The impact factor's Matthew effect: A natural experiment in bibliometrics. Journal of the American Society for Information Science and Technology, 61(2), 424-427.

Larivière, V., & Gingras, Y. (2011). Averages of ratios vs. ratios of averages: An empirical analysis of four levels of aggregation. Journal of Informetrics, 5(3), 392-399.

Larivière, V., & Macaluso, B. (2011). Improving the coverage of social science and humanities researchers' output: The case of the Érudit journal platform. Journal of the American Society for Information Science and Technology, 62(12), 2437-2442.

Larivière, V., Archambault, É., Gingras, Y., & Vignola-Gagné, É. (2006). The place of serials in referencing practices: Comparing natural sciences and engineering with social sciences and humanities. Journal of the American Society for Information Science and Technology, 57(8), 997-1004.

Larsen, P. O., & Von Ins, M. (2010). The rate of growth in scientific publication and the decline in coverage provided by Science Citation Index. Scientometrics, 84(3), 575-603.

Lasda Bergman, E. M. (2012). Finding citations to social work literature: The relative benefits of using Web of Science, Scopus, or Google Scholar. Journal of Academic Librarianship, 38(6), 370-379.

Latour, B. (1987). Science in action: How to follow scientists and engineers through society. Oxford: Oxford University Press.

Latour, B., & Woolgar, S. (1979). Laboratory life : the social construction of scientific facts. Beverly Hills : Sage Publications,.

Latour, B., & Woolgar, S. (1986). Laboratory life : the construction of scientific facts (Vol. 2nd, p. 294). Princeton, N.J.: Princeton University Press.

Laudel, G., & Gläser, J. (2006). Tensions between evaluations and communication practices. Journal of Higher Education Policy and Management, 28(3), 289-295.

Laudel, G., & Gläser, J. (2014). Beyond breakthrough research: Epistemic properties of research and their consequences for research funding. Research Policy, 43(7), 1204-1216.

Lee, C. J., Sugimoto, C. R., Zhang, G., & Cronin, B. (2013). Bias in peer review. Journal of the American Society for Information Science and Technology, 64(1), 2–17. doi:10.1002/asi.22784

Leisyte, L., & Dee, J. R. (2012). Understanding Academic Work in a Changing Institutional Environment. Higher Education: Handbook of Theory and Research, 27, 123-206.

Letierce, J., Passant, A., Breslin, J. G., & Decker, S. (2010). Using Twitter during an academic conference: The #iswc2009 use-case. Paper presented at the ICWSM 2010 – Proceedings of the 4th International AAAI Conference on Weblogs and Social Media, 279-282. http://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/viewFile/1523/1877

Letierce, J., Passant, A., Decker, S., & Breslin, J.G. (2010). Understanding how Twitter is used to spread scientific messages. In Proceedings of the Web Science Conference, Raleigh, NC, USA. Retrieved February 24, 2012 from http://journal.webscience.org/314/2/websci10_submission _79.pdf

Levitt, J. M., & Thelwall, M. (2011). A combined bibliometric indicator to predict article impact. Information Processing and Management, 47(2), 300-308.

Lewison, G. (2001). Evaluation of books as research outputs in history of medicine. Research Evaluation, 10(2), 89–95. doi:10.3152/147154401781777051

Leydesdorff, L. (2009). How are new citation-based journal indicators adding to the bibliometric toolbox? Journal of the American Society for Information Science and Technology, 60(7), 1327-1336.

Leydesdorff, L., & Bornmann, L. (2011a). How fractional counting of citations affects the impact factor: Normalization in terms of differences in citation potentials among fields of science. Journal of the American Society for Information Science and Technology, 62(2), 217-229.

Leydesdorff, L., & Bornmann, L. (2011b). Integrated impact indicators compared with impact factors: An alternative research design with policy implications. Journal of the American Society for Information Science and Technology, 62(11), 2133-2146.

Leydesdorff, L., & Bornmann, L. (in press). The operationalization of “fields” as WoS subject categories (WCs) in evaluative bibliometrics: The cases of “library and information science” and “science & technology studies”. Journal of the Association for Information Science and Technology.

Leydesdorff, L., & Opthof, T. (2010). Scopus's source normalized impact per paper (SNIP) versus a journal impact factor based on fractional counting of citations. Journal of the American Society for Information Science and Technology, 61(11), 2365-2369.

Leydesdorff, L., & Opthof, T. (2011). Remaining problems with the "new crown indicator" (MNCS) of the CWTS. Journal of Informetrics, 5(1), 224-225.

Leydesdorff, L., Bornmann, L., Mutz, R., & Opthof, T. (2011). Turning the tables on citation analysis one more time: Principles for comparing sets of documents. Journal of the American Society for Information Science and Technology, 62(7), 1370-1381.

Leydesdorff, L., Radicchi, F., Bornmann, L., Castellano, C., & Nooy, W. (2013b). Field-normalized impact factors (IFs): A comparison of rescaling and fractionally counted IFs. Journal of the American Society for Information Science and Technology, 64(11), 2299-2309.

Leydesdorff, L., Zhou, P., & Bornmann, L. (2013a). How can journal impact factors be normalized across fields of science? An assessment in terms of percentile ranks and fractional counts. Journal of the American Society for Information Science and Technology, 64(1), 96-107.

Li, J., Sanderson, M., Willett, P., Norris, M., & Oppenheim, C. (2010). Ranking of library and information science researchers: Comparison of data sources for correlating citation data, and expert judgments. Journal of Informetrics, 4(4), 554-563.

Li, X., & Thelwall, M. (2012). F1000, Mendeley and traditional bibliometric indicators. In: Proceedings of the 17th International Conference on Science and Technology Indicators. Montréal, Canada. pp. 451–551.

Li, X., Thelwall, M., & Giustini, D. (2012). Validating online reference managers for scholarly impact measurement. Scientometrics, 91(2), 461–471.

Li, Y., Radicchi, F., Castellano, C., & Ruiz-Castillo, J. (2013). Quantitative evaluation of alternative field normalization procedures. Journal of Informetrics,7(3), 746-755.

Lim, Y., Feng, D., & Cai, T. (2000). A web-based collaborative system for medical image analysis and diagnosis, ACM International Conference Proceeding Series; Selected papers from the Pan-Sydney workshop on Visualisation - Volume 2, pp. 93–95.

Lin, C. S., Huang, M. H., & Chen, D. Z. (2013). The influences of counting methods on university rankings based on paper count and citation count. Journal of Informetrics, 7(3), 611-621.

Lindsey, D. (1980). Production and citation measures in the sociology of science: The problem of multiple authorship. Social Studies of Science, 10(2), 145-162.

Linkova, M. (2014). Unable to resist: Researchers' responses to research assessment in the Czech Republic. Human Affairs, 24(1), 78-88.

Linmans, A. J. M. (2010). Why with bibliometrics the humanities does not need to be the weakest link. Indicators for research evaluation based on citations, library bindings and productivity measures. Scientometrics, 83(2), 337–354.

Lisée, C., Larivière, V., & Archambault, É. (2008). Conference proceedings as a source of scientific information: A bibliometric analysis. Journal of the American Society for Information Science and Technology, 59(11), 1776-1784.

Liu, X. Z., & Fang, H. (2012a). Fairly sharing the credit of multi-authored papers and its application in the modification of h-index and g-index. Scientometrics, 91(1), 37-49.

Liu, X. Z., & Fang, H. (2012b). Modifying h-index by allocating credit of multi-authored papers whose author names rank based on contribution. Journal of Informetrics, 6(4), 557-565.

Lock, S. (1994). Does Editorial Peer Review Work? Annals of Internal Medicine, 121(1), 60. doi:10.7326/0003-4819-121-1-199407010-00012

López-Cózar, E. D., Robinson-García, N., & Torres-Salinas, D. (2014). The Google Scholar experiment: How to index false papers and manipulate bibliometric indicators. Journal of the Association for Information Science and Technology, 65(3), 446-454.

López-Illescas, C., De Moya Anegón, F., & Moed, H. F. (2009). Comparing bibliometric country-by-country rankings derived from the Web of Science and Scopus: The effect of poorly cited journals in oncology. Journal of Information Science, 35(2), 244-256.

López-Illescas, C., De Moya-Anegón, F., & Moed, H. F. (2008). Coverage and citation impact of oncological journals in the Web of Science and Scopus. Journal of Informetrics, 2(4), 304-316.

Lozano, G. A., Larivière, V., & Gingras, Y. (2012). The weakening relationship between the impact factor and papers' citations in the digital age. Journal of the American Society for Information Science and Technology, 63(11), 2140-2145.

Lundberg, J. (2007). Lifting the crown—Citation z-score. Journal of informetrics, 1(2), 145-154.

Luukkonen, T. (2012). Conservatism and risk-taking in peer review: Emerging ERC practices. Research Evaluation 21(1): 48–60.

Luzón, M. J. (2007). Academic weblogs as tools for e-collaboration among researchers. In N. Kock (Ed.), Encyclopedia of e-collaboration (pp. 1–6). New York: Idea Group.

Luzón, M. J. (2009). Scholarly hyperwriting: The function of links in academic weblogs. Journal of the American Society for Information Science and Technology, 60(1), 75-89.

MacRoberts, M.H., & MacRoberts, B.R. (1989). Problems of citation analysis: A critical review. Journal of the American Society for Information Science 40(5): 342-349.

MacRoberts, M.H., & MacRoberts, B.R. (1996). Problems of citation analysis. Scientometrics, 36(3), 435-444.

Maflahi, N., & Thelwall, M. (in press). When are readers as good as citers for bibliometrics? Scopus vs. Mendeley for LIS journals. Journal of the Association for Information Science and Technology.

Mahdi, S., D'Este, P., & Neely, A. (2008). Citation counts: Are they good predictors of RAE scores? London: Advanced Institute of Management Research. http://digital.csic.es/bitstream/10261/11112/1/Citation%20Counts_%20Are%20they%20good%20predictors%20of%20rae%20scores.pdf

Mallard, G., Lamont, M., & Guetzkow, J. (2009). Fairness as Appropriateness: Negotiating Epistemological Differences in Peer Review. Science, Technology & Human Values, 34(5), 573–606. Retrieved from http://sth.sagepub.com/cgi/content/abstract/34/5/573

Manchikanti, L., Benyamin, R., Falco, F., Caraway,D., Datta, S., & Hirsch, J. (2012). Guidelines warfare over interventional techniques: is there a lack of discourse or straw man?. Pain Physician, 15, E1-E26.

Markpin, T., Boonradsamee, B., Ruksinsut, K., Yochai, W., Premkamolnetr, N., Ratchatahirun, P., & Sombatsompop, N. (2008). Article-count impact factor of materials science journals in SCI database. Scientometrics, 75(2), 251-261.

Marsh, H. W., Jayasinghe, U. W., & Bond, N. W. (2008). Improving the peer-review process for grant applications: reliability, validity, bias, and generalizability. The American Psychologist, 63(3), 160–8. doi:10.1037/0003-066X.63.3.160

Marshakova-Shaikevich, I. (1996). The standard impact factor as an evaluation tool of science fields and scientific journals. Scientometrics, 35(2), 283-290.

Martin, B. (1992). Scientific Fraud and the Power Structure of Science. Prometheus, 10(1), 83–98. doi:10.1080/08109029208629515

Martin, B. R. (2011). What can bibliometrics tell us about changes in the mode of knowledge production ? knowledge production ? Prometheus, 29(4), 455–479.

Martin, B. R., & Irvine, J. (1983). Assessing basic research: Some partial indicators of scientific progress in radio astronmy. Research Policy, 12(2), 61–90.

Martin, B., & Whitley, R. (2010). The UK Research Assessment Exercise: a case of regulatory capture. In: Whitley, R., Gläser, J., & Engwall, L. (eds.), Reconfiguring knowledge production: changing authority relationships in the sciences and their consequences for intellectual innovation (pp. 51-80). Oxford: Oxford University Press.

Marušić, A., Bošnjak, L., & Jerončić, A. (2011). A systematic review of research on the meaning, ethics and practices of authorship across scholarly disciplines. PLOS ONE, 6(9), e23477.

Mas-Bleda, A., Thelwall, M., Kousha, K., & Aguillo, I.F. (2014). Do highly cited researchers successfully use the Social Web? Scientometrics, 101(1), 337-356.

Mayr, P., & Walter, A. (2007). An exploratory study of Google Scholar. Online Information Review, 31(6), 814-830.

Mazloumian, A. (2012). Predicting scholars' scientific impact. PloS One, 7(11), e49246. doi:10.1371/journal.pone.0049246

McDonald, J. D. (2007). Understanding journal usage: A statistical analysis of citation and use. Journal of the American Society for Information Science and Technology, 58(1), 39-50.

McKay, S. (2012). Social Policy Excellence – Peer Review or Metrics? Analyzing the 2008 Research Assessment Exercise in Social Work and Social Policy and Administration. Social Policy & Administration, 46(5), 526–543. doi:10.1111/j.1467-9515.2011.00824.x

McKendrick, D. R. A., Cumming, G. P., & Lee, A. J. (2012). Increased use of Twitter at a medical conference: A report and a review of the educational opportunities. Journal of Medical Internet Research, 14(6).

Mcnay, I. (1998). The Research Assessment Exercise (RAE) and after: "You never know how it will all turn out." Perspectives: Policy and Practice in Higher Education, 2(1), 19–22.

Medoff, M. H. (2006). The efficiency of self-citations in economics. Scientometrics, 69(1), 69-84.

Meho, L. I., & Rogers, Y. (2008). Citation counting, citation ranking, and h-index of human-computer interaction researchers: A comparison of Scopus and Web of Science. Journal of the American Society for Information Science and Technology, 59(11), 1711-1726.

Meho, L. I., & Sonnenwald, D. H. (2000). Citation ranking versus peer evaluation of senior faculty research performance: A case study of Kurdish scholarship. Journal of the American Society for Information Science, 51(2), 123–138. doi:10.1002/(SICI)1097-4571(2000)51:2<123::AID-ASI4>3.0.CO;2-N

Meho, L. I., & Sugimoto, C. R. (2009). Assessing the scholarly impact of information studies: A tale of two citation databases—Scopus and Web of Science. Journal of the American Society for Information Science and Technology, 60(12), 2499-2508.

Meho, L. I., & Yang, K. (2007). Impact of data sources on citation counts and rankings of LIS faculty: Web of Science versus Scopus and Google Scholar. Journal of the American Society for Information Science and Technology, 58(13), 2105-2125.

Merton, R. K. (1968). The Matthew Effect in Science. Science, 159, 56–62.

Merton, R. K. (1973). The Normative Structure of Science (pp. 267–278). Chicago: University of Chicago Press.

Merton, R. K. (1988). The Matthew Effect in Science, II: Cumulative Advantage and the Symbolism of Intellectual Property. ISIS, 79, 606–623.

Metz, P., & Stemmer, J. (1996). A reputational study of academic publishers. College and Research Libraries, 57(3), 234-247.

Mewburn, I., & Thomson, P. (2013). Why do academics blog? An analysis of audiences, purposes and challenges. Studies in Higher Education, 38(8), 1105-1119.

Meyer, M. (2001). Patent citation analysis in a novel field of technology—An exploration of nano-science and nano-technology. Scientometrics, 51(1), 163–183.

Meyer, M. (2000a). What is special about patent citations? Differences between scientific and patent citations. Scientometrics, 49(1), 93-123.

Meyer, M. (2000b). Patent citations in a novel field of technology: What can they tell about interactions between emerging communities of science and technology? Scientometrics, 48(2), 151-178.

Meyer, M. (2002). Tracing knowledge flows in innovation systems. Scientometrics, 54(2), 193-212.

Michels, C., & Fu, J. Y. (2014). Systematic analysis of coverage and usage of conference proceedings in Web of Science. Scientometrics, 100(2), 307-327.

Michels, C., & Schmoch, U. (2012). The growth of science and database coverage. Scientometrics, 93(3), 831-846.

Mikki, S. (2010). Comparing Google Scholar and ISI Web of Science for earth sciences. Scientometrics, 82(2), 321-331.

Minasny B, Hartemink AE, McBratney A, Jang H. (2013) Citations and the h index of soil researchers and journals in the Web of Science, Scopus, and Google Scholar. PeerJ 1:e183 http://dx.doi.org/10.7717/peerj.183

Mingers, J., & Lipitakis, E. A. (2010). Counting the citations: A comparison of Web of Science and Google Scholar in the field of business and management. Scientometrics, 85(2), 613-625.

Mirowski, Ph. (2011). Science-Mart: Privatizing American Science. Cambridge, MA: Harvard University Press.

Mishra, J., Day, K., Littles, D., & Vandewalker, E. (2011). A content analysis of introductory courses in music education at NASM-accredited colleges and universities. Bulletin of the Council for Research in Music Education, (190), 7-19.

Moed, H. F. (2002). Measuring China's research performance using the Science Citation Index. Scientometrics, 53(3), 281-296.

Moed, H. F. (2005a). Citation analysis in research evaluation. Springer.

Moed, H. F. (2005b). Statistical relationships between downloads and citations at the level of individual documents within a single journal. Journal of the American Society for Information Science and Technology, 56(10), 1088–1097.

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(8), 575–583. doi:10.3152/030234207X255179

Moed, H. F. (2008). UK Research Assessment Exercises: Informed judgments on research quality or quantity?. Scientometrics, 74(1), 153-161.

Moed, H. F. (2010a). Measuring contextual citation impact of scientific journals. Journal of Informetrics, 4(3), 265-277.

Moed, H. F. (2010b). CWTS crown indicator measures citation impact of a research group's publication oeuvre. Journal of Informetrics, 4(3), 436-438.

Moed, H. F., & Glänzel, W. (2004). Handbook of quantitative science and technology research : the use of publication and patent statistics in studies of S&T systems. Dordrecht etc.: Kluwer Academic Publishers.

Moed, H. F., & Van Leeuwen, T. N. (1995). Improving the accuracy of Institute for Scientific Information's journal impact factors. Journal of the American Society for Information Science, 46(6), 461-467.

Moed, H. F., & Van Leeuwen, T. N. (1996). Impact factors can mislead. Nature, 381, 186.

Moed, H. F., & Visser, M. S. (2008). Appraisal of citation data sources. Centre for Science and Technology Studies, Leiden University. Retrieved November 29, 2014, from http://www.hefce.ac.uk/media/hefce/content/pubs/indirreports/2008/missing/Appraisal%20of%20Citation%20Data%20Sources.pdf.

Moed, H. F., Van Leeuwen, T. N., & Reedijk, J. (1998). A new classification system to describe the ageing of scientific journals and their impact factors. Journal of Documentation, 54(4), 387-419.

Moed, H. F., Van Leeuwen, T. N., & Reedijk, J. (1999). Towards appropriate indicators of journal impact. Scientometrics, 46(3), 575-589.

Mohammadi, E. (2014). Identifying the invisible impact of scholarly publications: a multi-disciplinary analysis using altmetrics (PhD Thesis). University of Wolverhampton.

Mohammadi, E., & Thelwall, M. (2013). Assessing non-standard article impact using F1000 labels. Scientometrics, 97(2), 383-395.

Mohammadi, E., & Thelwall, M. (2014). Mendeley readership altmetrics for the social sciences and humanities: Research evaluation and knowledge flows. Journal of the Association for Information Science and Technology, 65(8), 1627-1638.

Mohammadi, E., Thelwall, M., Haustein, S., & Larivière, V. (in press). Who reads research articles? An altmetrics analysis of Mendeley user categories. Journal of the Association for Information Science and Technology.

Moore, W. J., Newman, R. J., Sloane, P. J., & Steely, J. D. (2002). Productivity effects of research assessment exercises. University of Aberdeen, Department of Economics.

Mortensen, T., & Walker, J. (2002). Blogging thoughts: Personal publication as an online research tool. In Researching ICTs in Context, Edited by: Morrison, Andrew. 249-479. Oslo: Intermedia.http://possibleworlds.blogs.com/blogsperiment/files/Researching_ICTs_in_context-Ch11-Mortensen-Walker.pdf

Mryglod, O., Kenna, R., Holovatch, Y., & Berche, B. (2013). Comparison of a citation-based indicator and peer review for absolute and specific measures of research-group excellence. Scientometrics, 97(3), 767–777. doi:10.1007/s11192-013-1058-9

Mryglod, O., Kenna, R., Holovatch, Y., & Berche, B. (2014). Predicting Results of the Research Excellence Framework using departmental h-Index (pp. 1–13). Lviv, Ukraine. Retrieved from http://arxiv.org/abs/1411.1996v1

Mryglod, O., Kenna, R., Holovatch, Y., & Berche, B. (2015). Predicting Results of the Research Excellence Framework using Departmental h -Index – Revisited (No. arXiv:1501.07857v1). Lviv, Ukraine.

Musselin, C. (2013). How peer review empowers the academic profession and university managers: Changes in relationships between the state, universities and the professoriate. Research Policy, 42(5), 1165-1173.

Mutz, R., & Daniel, H. D. (2012a). The generalized propensity score methodology for estimating unbiased journal impact factors. Scientometrics,92(2), 377-390.

Mutz, R., & Daniel, H. D. (2012b). Skewed citation distributions and bias factors: Solutions to two core problems with the journal impact factor. Journal of Informetrics, 6(2), 169-176.

Mutz, R., Bornmann, L., & Daniel, H.-D. (2012a). Does Gender Matter in Grant Peer Review? Zeitschrift Für Psychologie, 220(2), 121–129. doi:10.1027/2151-2604/a000103

Mutz, R., Bornmann, L., & Daniel, H.-D. (2012b). Heterogeneity of inter-rater reliabilities of grant peer reviews and its determinants: a general estimating equations approach. PloS One, 7(10), e48509. doi:10.1371/journal.pone.0048509

Nederhof, A. J. (2006). Bibliometric monitoring of research performance in the social sciences and the humanities: A review. Scientometrics, 66(1), 81-100.

Nederhof, A. J., & Van Raan, A. F. J. (1993). A bibliometric analysis of six economics research groups: A comparison with peer review. Research Policy, 22(4), 353–368. doi:10.1016/0048-7333(93)90005-3

Nederhof, T. (1988). The validity and reliability of evaluation of scholarly performance. In A. van Raan (Ed.), Handbook of Quantitative Studies of Science and Technology (pp. 193–228). Burlington : Elsevier Science,.

Nedeva, M., Boden, R., & Nugroho, Y. (2012). Rank and File: managing individual performance in university research. Higher Education Policy, 25(3), 335-360.

Neill, A., Cronin, J. J., Brannigan, D., O'Sullivan, R., & Cadogan, M. (2014). The impact of social media on a major international emergency medicine conference. Emergency Medicine Journal, 31(5), 401-404.

Neuhaus, C., & Daniel, H. D. (2009). A new reference standard for citation analysis in chemistry and related fields based on the sections of Chemical Abstracts. Scientometrics, 78(2), 219-229.

Neylon, C., & Wu. S. (2009). Article-level metrics and the evolution of scientific impact. PLoS Biol 7(11). DOI: 10.1371/journal.pbio.1000242

Nicholas, D., Clark, D., Rowlands, I., & Jamali, H. R. (2009). Online use and information seeking behaviour: Institutional and subject comparisons of UK researchers. Journal of Information Science, 35(6), 660-676.

Nicolaisen, J. (2002). The scholarliness of published peer reviews: A bibliometric study of book reviews in selected social science fields. Research Evaluation, 11(3), 129-140.

Nicolaisen, J. (2007). Citation Analysis. In Annual Review of Information Science and Technology, 41 (pp. 609–642). New York etc.: Interscience Publishers.

Nicolaisen, J., & Frandsen, T. F. (2008). The reference return ratio. Journal of Informetrics, 2(2), 128-135.

Nielsen, F. A. (2007). Scientific citations in Wikipedia. First Monday, 12(8). http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/1997/1872

Nightingale, P., & Scott, A. (2007). Peer review and the relevance gap: ten suggestions for policy-makers. Science and Public Policy, 34(8), 543–553. doi:10.3152/030234207X254396

Noijons, E., & Wouters, P. (2014). Report on the 19th International Conference on Science and Technology Indicators. ISSI Newsletter, 10(4), 69–70.

Norris, M., & Oppenheim, C. (2003). Citation counts and the Research Assessment Exercise V. Journal of Documentation, 59(6), 709–730. doi:10.1108/00220410310698734

Norris, M., & Oppenheim, C. (2007). Comparing alternatives to the Web of Science for coverage of the social sciences' literature. Journal of Informetrics, 1(2), 161-169.

Norris, M., & Oppenheim, C. (2010a). Peer review and the h-index: Two studies. Journal of Informetrics, 4(3), 221–232. doi:10.1016/j.joi.2009.11.001

Norris, M., & Oppenheim, C. (2010b). The h-index: A broad review of a new bibliometric indicator. Journal of Documentation, 66(5), 681-705.

Nowotny, H., Scott, P., & Gibbons, M. (2001). Re-thinking Science: Knowledge and the Public in an Age of Uncertainty. Cambridge (UK): Polity Press.

Ochsner, M., Hug, S. E., & Daniel, H. (2012). Indicators for Research Quality for Evaluation of Humanities Research : Opportunities and Limitations Quality criteria for research in the humanities Collecting indicators for research in the humanities The measurement of research quality in the humanitie. Bibliometrie – Praxis und Forschung, UND FORSCHUNG, 1(4), 1–17.

Ochsner, M., Hug, S. E., & Daniel, H.-D. (2014). Setting the stage for the assessment of research quality in the humanities. Consolidating the results of four empirical studies. Zeitschrift Für Erziehungswissenschaft, 17(S6), 111–132. doi:10.1007/s11618-014-0576-4

Olbrecht, M., & Bornmann, L. (2010). Panel peer review of grant applications: what do we know from research in social psychology on judgment and decision-making in groups? Research Evaluation, 19(4), 293–304. doi:10.3152/095820210X12809191250762

Oppenheim, C. (1995). The correlation between citation counts and the 1992 research assessment exercise ratings for british library and information science university departments. Journal of Documentation, 51(1), 18–27. doi:10.1108/eb026940

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.

Oppenheim, C. (2000). Do patent citations count? In: B. Cronin & H B. Atkins (Eds.), The web of knowledge: A festschrift in honor of Eugene Garfield (pp. 405-432). Metford, NJ. Information Today Inc.

Opthof, T., & Leydesdorff, L. (2010). Caveats for the journal and field normalizations in the CWTS (“Leiden”) evaluations of research performance. Journal of Informetrics, 4(3), 423-430.

Opthof, T., Furstner, F., Geer, M. Van, & Coronel, R. (2000). Regrets or no regrets ? No regrets ! The fate of rejected manuscripts. Cardiavascular Research, 45, 255–258.

Orduña-Malea, E., & Delgado López-Cózar, E. (2014). Google Scholar metrics evolution: An analysis according to languages. Scientometrics, 98(3), 2353-2367.

Ortega, J. L., & Aguillo, I. F. (2012). Science is all in the eye of the beholder: Keyword maps in Google Scholar citations. Journal of the American Society for Information Science and Technology, 63(12), 2370-2377.

Ortega, J. L., & Aguillo, I. F. (2014). Microsoft academic search and Google Scholar citations: Comparative analysis of author profiles. Journal of the Association for Information Science and Technology, 65(6), 1149-1156.

Ossenblok, T. L., Engels, T. C., & Sivertsen, G. (2012). The representation of the social sciences and humanities in the Web of Science—a comparison of publication patterns and incentive structures in Flanders and Norway (2005–9). Research Evaluation, 21(4), 280-290.

Overbeeke, J., Godlee, F., & Jefferson, T. (1999). The state of evidence: what we know and what we don't know about journal peer review. London: BMJ Books.

Paltoglou, G., & Thelwall, M. (2012). Twitter, MySpace, Digg: Unsupervised sentiment analysis in social media. ACM Transactions on Intelligent Systems and Technology, 3(4)

Panaretos, J., & Malesios, C. (2009). Assessing scientific research performance and impact with single indices. Scientometrics, 81(3), 635-670.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 2(1-2), 1-135.

Paradeise, C., & Thoenig, J. C. (2013). Academic institutions in search of quality: Local orders and global standards. Organization studies, 34(2), 189-218.

Penner, O., Petersen, A. M., Pan, R. K., & Fortunato, S. (2013). Commentary: The case for caution in predicting scientists' future impact. Physics Today, 66(4), 8. doi:10.1063/PT.3.1928

Persson, O., Glänzel, W., & Danell, R. (2004). Inflationary bibliometric values: The role of scientific collaboration and the need for relative indicators in evaluative studies. Scientometrics, 60(3), 421-432.

Pezzoni, M., Sterzi, V., & Lissoni, F. (2012). Career progress in centralized academic systems: Social capital and institutions in France and Italy. Research Policy, 41(4), 704–719. doi:10.1016/j.respol.2011.12.009

Pfeffer, J., & Salancik, G. R. (2003). The external control of organizations: A resource dependence perspective. Stanford, CA: Stanford University Press.

Phillips, M. (2012). Research universities and research assessment.

Pieterse, A. L., Evans, S. A., Risner-Butner, A., Collins, N. M., & Mason, L. B. (2009). Multicultural competence and social justice training in counseling psychology and counselor education: A review and analysis of a sample of multicultural course syllabi. Counseling Psychologist, 37(1), 93-115.

Pinski, G., & Narin, F. (1976). Citation influence for journal aggregates of scientific publications: Theory, with application to the literature of physics. Information Processing and Management, 12(5), 297-312.

Piwowar, H. A., Day, R. S., & Fridsma, D. B. (2007). Sharing detailed research data is associated with increased citation rate. PLoS ONE, 2(3): e308. doi:10.1371/journal.pone.0000308

Plomp, R. (1990). The significance of the number of highly cited papers as an indicator of scientific prolificacy. Scientometrics, 19(3), 185-197.

Plomp, R. (1994). The highly cited papers of professors as an indicator of a research group's scientific performance. Scientometrics, 29(3), 377-393.

Pontille, D., & Torny, D. (2010). The controversial policies of journal ratings: Evaluating social sciences and humanities. Research Evaluation, 19(5), 347-360.

Pontille, D., & Torny, D. (2014). The Blind Shall See! The Question of Anonymity in Journal Peer Review. Ada: A Journal of Gender, New Media, and Technology, ("4"), 1–15. doi:10.7264/N3542KVW

Porter, Th. (1995). Trust in numbers: the pursuit of objectivity in science and public life. Princeton, NY: Princeton University Press.

Power, M. (1997). The audit society: Rituals of verification. Oxford: Oxford University Press.

Priem, J., & Costello, K. L. (2010), How and why scholars cite on Twitter. In proceedings of the American Society for Information Science and Technology, (47)1, 1-4.

Priem, J., & Hemminger, B. M. (2010). Scientometrics 2.0: Toward new metrics of scholarly impact on the social web. First Monday, 15(7). http://firstmonday.org/ojs/index.php/fm/article/view/2874/2570

Priem, J., Piwowar, H., & Hemminger, B. (2012). Altmetrics in the wild: Using social media to explore scholarly impact. Retrieved from http://arXiv.org/html/1203.4745v1

Priem, J., Taraborelli, D., Groth, P., & Neylon, C. (2010). Altmetrics: A manifesto. Retrieved from http://altmetrics.org/manifesto/

Pudovkin, A. I., & Garfield, E. (2004). Rank-normalized impact factor: A way to compare journal performance across subject categories. Proceedings of the American Society for Information Science and Technology, 41(1), 507-515.

Pudovkin, A. I., & Garfield, E. (2009). Percentile rank and author superiority indexes for evaluating individual journal articles and the author's overall citation performance. COLLNET Journal of Scientometrics and Information Management,3(2), 3-10.

Puschmann, C., & Mahrt, M. (2012). Scholarly blogging: A new form of publishing or science journalism 2.0? In A. Tokar, M. Beurskens, S. Keuneke, M. Mahrt, I. Peters, C. Puschmann, K. Weller, & T. van Treeck (eds.), Science and the Internet (pp. 171-182). Düsseldorf: Düsseldorf University Press.

Radder, H. (ed.) (2010). The commodification of academic research: science and the modern university. Pittsburgh: University of Pittsburgh Press.

Radicchi, F., & Castellano, C. (2011). Rescaling citations of publications in physics. Physical Review E, 83(4), 046116.

Radicchi, F., & Castellano, C. (2012a). Testing the fairness of citation indicators for comparison across scientific domains: The case of fractional citation counts. Journal of Informetrics, 6(1), 121-130.

Radicchi, F., & Castellano, C. (2012b). A reverse engineering approach to the suppression of citation biases reveals universal properties of citation distributions. PLOS ONE, 7(3), e33833.

Radicchi, F., Fortunato, S., & Castellano, C. (2008). Universality of citation distributions: Toward an objective measure of scientific impact. Proceedings of the National Academy of Sciences of the United States of America, 105(45), 17268-17272.

Rafols, I., Leydesdorff, L., O'Hare, A., Nightingale, P., & Stirling, A. (2012). How journal rankings can suppress interdisciplinary research: A comparison between innovation studies and business & management. Research Policy, 41(7), 1262-1282.

Ragone, A., Mirylenka, K., Casati, F., & Marchese, M. (2013). On peer review in computer science: analysis of its effectiveness and suggestions for improvement. Scientometrics, 97(2), 317–356. doi:10.1007/s11192-013-1002-z

Rasmussen, P. G., & Andersen, J.P. (2013). Altmetrics: An alternate perspective on research evaluation. Sciecom Info, 9(2). http://journals.lub.lu.se/index.php/sciecominfo/article/view/7292/6102

Reale, E., & Seeber, M. (2013). Instruments as empirical evidence for the analysis of Higher Education policies. Higher Education, 65(1), 135-151.

Rebora, G., & Turri, M. (2013). The UK and Italian research assessment exercises face to face. Research Policy, 42(9), 1657-1666.

REF (2011). Decisions on assessing research impact. Retrieved from http://www.ref.ac.uk/media/ref/content/pub/decisionsonassessingresearchimpact/01_11.pdf

REF (2012). Panel criteria and working methods. http://www.ref.ac.uk/media/ref/content/pub/panelcriteriaandworkingmethods/01_12.doc

Reinhart, M. (2009). Peer review of grant applications in biology and medicine. Reliability, fairness, and validity. Scientometrics, 81(3), 789–809. doi:10.1007/s11192-008-2220-7

Reinhart, M. (2010). Peer review practices: a content analysis of external reviews in science funding. Research Evaluation, 19(5), 317–331. doi:10.3152/095820210X12809191250843

Rinia, E., van Leeuwen, T., van Vuren, H., & van Raan, A. F. (1998). Comparative analysis of a set of bibliometric indicators and central peer review criteria. Research Policy, 27(1), 95–107. doi:10.1016/S0048-7333(98)00026-2

Rinia, E. J., De Lange, C., & Moed, H. F. (1993). Measuring national output in physics: Delimitation problems. Scientometrics, 28(1), 89-110.

Ritterbusha, J. (2007). Supporting library research with LibX and Zotero: Two open source Firefox extensions. Journal of Web Librarianship, 1(3), 111-122.

Rons, N. (2012). Partition-based field normalization: An approach to highly specialized publication records. Journal of Informetrics, 6(1), 1-10.

Ross, C., Terras, M., Warwick, C., & Welsh, A. (2011). Enabled backchannel: Conference Twitter use by digital humanists. Journal of Documentation, 67(2), 214-237.

Rossi, B., Russo, B., & Succi, G. (2010). Download patterns and releases in open source software projects: A perfect symbiosis? 6th International IFIP WG 2.13 Conference on Open Source Systems, Notre Dame, IN; United States; 30 May-2 June. Volume 319, pp. 252-267. http://www.inf.unibz.it/~gsucci/publications/images/DownloadPatternsandReleasesinOpenSourceSoftwareProjectsaPerfectSymbiosis.pdf

Rousseau, R. (1997). Sitations: An exploratory study. Cybermetrics, 1(1), Retrieved November 14, 2001, from http://www.cindoc.csic.es/cybermetrics/articles/v2i1p2.html

Rousseau, R. (2002). Journal evaluation: Technical and practical issues. Library Trends, 50(3), 418-439.

Rousseau, R. (2005). Median and percentile impact factors: A set of new indicators. Scientometrics, 63(3), 431-441.

Rousseau, R., & Fred, Y. Y. (2013). A multi-metric approach for research evaluation. Chinese Science Bulletin, 58(26), 3288-3290.

Rousseau, S., & Rousseau, R. (2012). Interactions between journal attributes and authors' willingness to wait for editorial decisions. Journal of the American Society for Information Science and Technology, 63(6), 1213-1225.

Rowlands, I., & Nicholas, D. (2007). The missing link: Journal usage metrics. Aslib Proceedings: New Information Perspectives, 59(3), 222-228.

Ruiz-Castillo, J., & Waltman, L. (in press). Field-normalized citation impact indicators using algorithmically constructed classification systems of science. Journal of Informetrics.

Rushforth, A. and De Rijcke, S. Accounting for Impact? The Journal Impact Factor and the Making of Biomedical Research in the Netherlands. (Manuscript submitted for publication.)

Sá, C. M., Kretz, A., & Sigurdson, K. (2013). Accountability, performance assessment, and evaluation: Policy pressures and responses from research councils. Research Evaluation, 22(2), 105-117.

Sauder, M., & Espeland, W. N. (2009). The discipline of rankings: Tight coupling and organizational change. American Sociological Review, 74(1), 63-82.

Schimank, U. (2005). “New Public Management” and the Academic Profession: Reflections on the German Situation. Minerva 43(4): 361–376.

Schloegl, C., & Gorraiz, J. (2010). Comparison of citation and usage indicators: The case of oncology journals. Scientometrics, 82(3), 567-580.

Schneider, J. W. (2009). An outline of the bibliometric indicator used for performance-based funding of research institutions in Norway. European Political Science, 8(3), 364-378.

Schonfeld, R.C., & Housewright, R. (2010). Faculty survey 2009: Key strategic insights for libraries, publishers, and societies. Ithaka S+R, New York, NY. http://www.sr.ithaka.org/sites/default/files/reports/Faculty_Study_2009.pdf

Schreiber, M. (2007). Self-citation corrections for the Hirsch index. EPL, 78(3), 30002.

Schreiber, M. (2008a). The influence of self-citation corrections on Egghe’s g index. Scientometrics, 76(1), 187-200.

Schreiber, M. (2008b). A modification of the h-index: The hm-index accounts for multi-authored manuscripts. Journal of Informetrics, 2(3), 211-216.

Schreiber, M. (2008c). To share the fame in a fair way, hm modifies h for multi-authored manuscripts. New Journal of Physics, 10(4), 040201.

Schreiber, M. (2009a). A case study of the modified Hirsch index hm accounting for multiple coauthors. Journal of the American Society for Information Science and Technology, 60(6), 1274-1282.

Schreiber, M. (2009b). Fractionalized counting of publications for the g-index. Journal of the American Society for Information Science and Technology, 60(10), 2145-2150.

Schreiber, M. (2010a). How to modify the g-index for multi-authored manuscripts. Journal of Informetrics, 4(1), 42-54.

Schreiber, M. (2010b). A case study of the modified g index: Counting multi-author publications fractionally. Journal of Informetrics, 4(4), 636-643.

Schreiber, M. (2013). How much do different ways of calculating percentiles influence the derived performance indicators? A case study. Scientometrics,97(3), 821-829.

Schröder, K., & Lüthen, H. (2009). Astrophotography. In: In: G. D. Roth, (Eds), Handbook of practical astronomy. Berlin: Springer, pp. 133-173.

Schubert, A., & Braun, T. (1993). Reference standards for citation based assessments. Scientometrics, 26(1), 21-35.

Schubert, A., & Braun, T. (1996). Cross-field normalization of scientometric indicators. Scientometrics, 36(3), 311-324.

Schubert, A., Glänzel, W., & Braun, T. (1989). Scientometric datafiles. A comprehensive set of indicators on 2649 journals and 96 countries in all major science fields and subfields 1981–1985. Scientometrics, 16(1), 3-478.

Schubert, A., Glänzel, W., & Thijs, B. (2006). The weight of author self-citations. A fractional approach to self-citation counting. Scientometrics, 67(3), 503-514.

Seglen, P. O. (1992). The skewness of science. Journal of the American Society for Information Science, 43(9), 628-638.

Seglen, P. O. (1994). Causal relationship between article citedness and journal impact. Journal of the American Society for Information Science, 45(1), 1-11.

Seglen, P. O. (1997). Why the impact factor of journals should not be used for evaluating research. British Medical Journal, 314(7079), 498-502.

Sen, B. K. (1992). Normalised impact factor. Journal of Documentation, 48(3), 318-325.

Serenko, A., & Dohan, M. (2011). Comparing the expert survey and citation impact journal ranking methods: Example from the field of Artificial Intelligence. Journal of Informetrics, 5(4), 629–648. doi:10.1016/j.joi.2011.06.002

Shapin, S. (1998). The Scientific Revolution. University Of Chicago Press. Retrieved from http://www.amazon.com/dp/0226750213

Shaw, D. (1991). An analysis of the relationship between book reviews and fiction holdings in OCLC. Library and Information Science Research, 13(2), 147-154.

Shema, H., Bar-Ilan, J., & Thelwall, M. (2012). Research blogs and the discussion of scholarly information. PLoS ONE, 7(5).

Shema, H., Bar-Ilan, J., & Thelwall, M. (2014). Do blog citations correlate with a higher number of future citations? Research blogs as a potential source for alternative metrics. Journal of the Association for Information Science and Technology, 65(5), 1018-1027.

Shen, H. W., & Barabási, A. L. (2014). Collective credit allocation in science. Proceedings of the National Academy of Sciences of the United States of America, 111(34), 12325-12330.

Shore, C. (2008). Audit culture and Illiberal governance Universities and the politics of accountability. Anthropological Theory, 8(3), 278-298.

Shore, C. (2010). Beyond the multiversity: neoliberalism and the rise of the schizophrenic university. Social Anthropology, 18(1), 15-29.

Shuai, X., Pepe, A., & Bollen, J. (2012). How the scientific community reacts to newly submitted preprints: Article downloads, Twitter mentions, and citations. PLoS ONE, 7(11)

Sigogneau, A. (2000). An analysis of document types published in journals related to physics: Proceeding papers recorded in the Science Citation Index database. Scientometrics, 47(3), 589-604.

Simons, K. (2008). The misused impact factor. Science, 322(5899), 165-165.

Sirtes, D. (2012). Finding the Easter eggs hidden by oneself: Why Radicchi and Castellano's (2012) fairness test for citation indicators is not fair. Journal of Informetrics, 6(3), 448-450.

Sivertsen, G., & Larsen, B. (2012). Comprehensive bibliographic coverage of the social sciences and humanities in a citation index: An empirical analysis of the potential. Scientometrics, 91(2), 567-575.

Smith, A., & Thelwall, M. (2002). Web impact factors for Australasian universities. Scientometrics, 54(3), 363-380.

Smith, A.G. (1999). A tale of two Web spaces: Comparing sites using Web impact factors. Journal of Documentation, 55(5), 577-592.

Smith, R. (2006). Peer review: a flawed process at the heart of science and journals. Journal of the Royal Society of Medicine, 99(4), 178–182. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1420798/

Sombatsompop, N., & Markpin, T. (2005). Making an equality of ISI impact factors for different subject fields. Journal of the American Society for Information Science and Technology, 56(7), 676-683.

Sombatsompop, N., Markpin, T., & Premkamolnetr, N. (2004). A modified method for calculating the impact factors of journals in ISI Journal Citation Reports: Polymer Science category in 1997–2001. Scientometrics, 60(2), 217-235.

Sousa, S. B., & Brennan, J. L. (2014). The UK Research Excellence Framework and the Transformation of Research Production. In: Musselin, C., & Teixeira, P. (eds.), Reforming Higher Education (pp. 65-80). Dordrecht: Springer.

Stachowiak, B. (2014). The presence of Polish academics on social networking websites for academics, using the example of employees of Nicolaus Copernicus University, Universal Journal of

Educational Research 2(1), 64-68. [On Beall's list of predatory open access publishers: http://scholarlyoa.com/publishers/]

Stallings, J., Vance, E., Yang, J., Vannier, M. W., Liang, J., Pang, L., ... & Wang, G. (2013). Determining scientific impact using a collaboration index. Proceedings of the National Academy of Sciences of the United States of America, 110(24), 9680-9685.

Stankus, T., & Rice, B. (1982). Handle with care: Use and citation data for science journal management. Collection Management, 4(1-2), 95-110.

Steen, R. G. (2011). Retractions in the scientific literature: do authors deliberately commit research fraud?. Journal of Medical Ethics, 37(2), 113-117.

Stephan, P.E. (2012). How economics shapes science, Cambridge, MA: Harvard University Press.

Stern, D. I. (2014). High-ranked social science journal articles can be identified from early citation information. PLOS ONE, 9(11), e112520.

Stöckelová, T. (2014). Power at the Interfaces: The Contested Orderings of Academic Presents and Futures in a Social Science Department. Higher Education Policy, 27(4), 435-451.

Strathern, M. (ed.) (2000). Audit Cultures: Anthropological Studies in Accountability, Ethics and the Academy. London: Routledge.

Sud, P., & Thelwall, M. (2014a). Evaluating altmetrics. Scientometrics, 98(2), 1131-1143.

Sud, P., & Thelwall, M. (2014b). Linked title mentions: A new automated link search candidate. Scientometrics. DOI 10.1007/s11192-014-1374-8

Sugimoto, C. R., & Cronin, B. (2013). Citation gamesmanship: testing for evidence of ego bias in peer review. Scientometrics, 95(3), 851-862.

Sugimoto, C. R., & Thelwall, M. (2013). Scholars on soap boxes: Science communication and dissemination in TED videos. Journal of the American Society for Information Science and Technology, 64(4), 663-674.

Suls, J., & Martin, R. (2009). The Air We Breathe: A Critical Look at Practices and Alternatives in the Peer-Review Process. Perspectives on Psychological Science, 4(1), 40–50. doi:10.1111/j.1745-6924.2009.01105.x

Taraborelli, D. (2008). Soft peer review: Social software and distributed scientific evaluation, Proceedings of the Eighth International Conference on the Design of Cooperative Systems, Carry–Le–Rouet, 20–23 May. http://nitens.org/docs/spr_coop08.pdf

Taylor, J. (1994). Measuring Research Performance in Business and Management Studies in the United Kingdom: The 1992 Research Assessment Exercise. British Journal of Management, 5, 275–288.

Taylor, J. (2011). The Assessment of Research Quality in UK Universities: Peer Review or Metrics? British Journal of Management, 22(2), 202–217. doi:10.1111/j.1467-8551.2010.00722.x

Taylor, J., & Walker, I. (2009). Peer assessment of research: How many publications per staff? Lancaster University Management School, Working Paper 2009/035. http://eprints.lancs.ac.uk/31757/1/006236.pdf

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, AU., Wu, L., et al. (2011). Data sharing by scientists: practices and perceptions. PLoS ONE 6(6): e21101. doi:10.1371/journal.pone.0021101

Thelwall, M. (2001). Extracting macroscopic information from web links, Journal of the American Society for Information Science and Technology, 52(13), 1157-1168.

Thelwall, M. (2012). Journal impact evaluation: A webometric perspective. Scientometrics, 92(2), 429-441.

Thelwall, M., & Maflahi, N. (in press). Guideline references and academic citations as evidence of the clinical value of health research. Journal of the Association for Information Science and Tec,hnology. http://www.scit.wlv.ac.uk/~cm1993/papers/GuidelineMetricsPreprint.pdf

Thelwall, M., & Harries, G. (2003). The connection between the research of a university and counts of links to its web pages: An investigation based upon a classification of the relationships of pages to the research of the host university. Journal of the American Society for Information Science and Technology, 54(7), 594-602.

Thelwall, M., & Kousha, K. (2008). Online presentations as a source of scientific impact?: An analysis of PowerPoint files citing academic journals. Journal of the American Society for Information Science & Technology, 59(5), 805-815.

Thelwall, M., & Kousha, K. (2014a). Academia.edu: Social network or academic network?. Journal of the Association for Information Science and Technology, 65 (4), 721–731.

Thelwall, M., & Kousha, K. (2014b). ResearchGate: Disseminating, communicating, and measuring scholarship? Journal of the Association for Information Science and Technology.
doi: 10.1002/asi.23236

Thelwall, M., Buckley, K., & Paltoglou, G. (2011). Sentiment in Twitter events. Journal of the American Society for Information Science and Technology, 62(2), 406-418.

Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., & Kappas, A. (2010). Sentiment strength detection in short informal text. Journal of the American Society for Information Science and Technology, 61(12), 2544-2558.

Thelwall, M., Haustein, S., Larivière, V., & Sugimoto, C. (2013). Do altmetrics work? Twitter and ten other candidates. PLOS ONE, 8(5), e64841. doi:10.1371/journal.pone.0064841

Thelwall, M., Sud, P., & Vis, F. (2012). Commenting on YouTube videos: From Guatemalan rock to El Big Bang. Journal of the American Society for Information Science and Technology, 63(3), 616–629.

Thelwall, M., Tsou, A., Weingart, S., Holmberg, K., & Haustein, S (2013) Tweeting links to academic articles. Cybermetrics, 17(1). http://cybermetrics.cindoc.csic.es/articles/v17i1p1.html.

Thijs, B., & Glänzel, W. (2006). The influence of author self-citations on bibliometric meso-indicators. The case of European universities. Scientometrics, 66(1), 71-80.

Thomas, P. R., & Watkins, D. S. (1998). Institutional research rankings via bibliometric analysis and direct peer review: A comparative case study with policy implications. Scientometrics, 41(3), 335–355. doi:10.1007/BF02459050

Thompson, J. W. (2002). The death of the scholarly monograph in the humanities? Citation patterns in literary scholarship. Libri, 52(3), 121-136.

Tijssen, R. J., Visser, M. S., & Van Leeuwen, T. N. (2002). Benchmarking international scientific excellence: Are highly cited research papers an appropriate frame of reference? Scientometrics, 54(3), 381-397.

Tol, R. S. (2011). Credit where credit's due: Accounting for co-authorship in citation counts. Scientometrics, 89(1), 291-299.

Torres-Salinas, D., & Moed, H. F. (2009). Library catalog analysis as a tool in studies of social sciences and humanities: An exploratory study of published book titles in economics. Journal of Informetrics, 3(1), 9-26.

Torres-Salinas, D., Lopez-Cózar, E. D., & Jiménez-Contreras, E. (2009). Ranking of departments and researchers within a university using two different databases: Web of Science versus Scopus. Scientometrics, 80(3), 761-774.

Torres-Salinas, D., Robinson-García, N., Jiménez-Contreras, E., & Delgado López-Cózar, E. (2012). Towards a 'Book Publishers Citation Reports'. First approach using the 'Book Citation Index'. Revista Española de Documentación Científica, 35(4), 615-620.

Torres-Salinas, D., Rodríguez-Sánchez, R., Robinson-García, N., Fdez-Valdivia, J., & García, J. A. (2013). Mapping citation patterns of book chapters in the Book Citation Index. Journal of Informetrics, 7(2), 412-424.

Trueba, F. J., & Guerrero, H. (2004). A robust formula to credit authors for their publications. Scientometrics, 60(2), 181-204.

Tsay, M. (1998). The relationship between journal use in a medical library and citation use. Bulletin of the Medical Library Association, 86(1), 31-39.

Van Arensbergen, P. (2014). Talent proof : selection processes in research funding and careers. The Hague, Netherlands: Rathenau Institute. Retrieved from http://www.worldcat.org/title/talent-proof-selection-processes-in-research-funding-and-careers/oclc/890766139&referer=brief_results

Van Dalen, H. P., & Henkens, K. (2012). Intended and unintended consequences of a publish-or-perish culture: A worldwide survey. Journal of the American Society for Information Science and Technology, 63(7), 1282-1293.

Van den Besselaar, P. (2012). Selection committee membership: Service or self-service. Journal of Informetrics, 6(4), 580–585. doi:10.1016/j.joi.2012.05.003

Van den Besselaar, P., & Leydesdorff, L. (2009). Past performance, peer review and project selection: a case study in the social and behavioral sciences. Research Evaluation, 18(4), 273–288. doi:10.3152/095820209X475360

Van der Meulen, B. (2007). Interfering Governance and Emerging Centres of Control: University research evaluation in the Netherlands. In: Whitley, R., & Gläser, J. (eds.), The Changing Governance of the Sciences (pp. 191-203). Dordrecht: Springer.

Van Eck, N. J., Waltman, L., Van Raan, A. F., Klautz, R. J., & Peul, W. C. (2013). Citation analysis may severely underestimate the impact of clinical research as compared to basic research. PLOS ONE, 8(4), e62395.

Van Hooydonk, G. (1997). Fractional counting of multiauthored publications: Consequences for the impact of authors. Journal of the American Society for Information Science, 48(10), 944-945.

Van Leeuwen, T. N., & Calero-Medina, C. (2012). Redefining the field of economics: Improving field normalization for the application of bibliometric techniques in the field of economics. Research Evaluation, 21(1), 61-70.

Van Leeuwen, T. N., & Moed, H. F. (2002). Development and application of journal impact measures in the Dutch science system. Scientometrics, 53(2), 249-266.

Van Leeuwen, T. N., & Moed, H. F. (2005). Characteristics of journal impact factors: the effects of uncitedness and citation distribution on the understanding of journal impact factors. Scientometrics, 63(2), 357-371.

Van Leeuwen, T. N., & Moed, H. F. (2012). Funding decisions, peer review, and scientific excellence in physical sciences, chemistry, and geosciences. Research Evaluation, 21(3), 189–198. doi:10.1093/reseval/rvs009

Van Leeuwen, T. N., Moed, H. F., Tijssen, R. J., Visser, M. S., & Van Raan, A. F. (2001). Language biases in the coverage of the Science Citation Index and its consequences for international comparisons of national research performance. Scientometrics, 51(1), 335-346.

Van Leeuwen, T. N., Van der Wurff, L. J., & De Craen, A. J. M. (2007). Classification of “research letters” in general medical journals and its consequences in bibliometric research evaluation processes. Research Evaluation, 16(1), 59-63.

Van Leeuwen, T. N., Visser, M. S., Moed, H. F., Nederhof, T. J., & Van Raan, A. F. (2003). The Holy Grail of science policy: Exploring and combining bibliometric tools in search of scientific excellence. Scientometrics, 57(2), 257-280.

Van Leeuwen, T., Costas, R., Calero-Medina, C., & Visser, M. (2013). The role of editorial material in bibliometric research performance assessments. Scientometrics, 95(2), 817-828.

Van Noorden, R. (2010). Metrics: A profusion of measures. Nature, 465(7300), 864–866.

Van Raan, A. (Ed.). (1988). Handbook of Quantitative Studies of Science and Technology. Amsterdam: Elsevier Science Publishers.

Van Raan, A. F. J. (1996). Advanced bibliometric methods as quantitative core of peer review based evaluation and foresight exercises. Scientometrics, 36(3), 397–420.

Van Raan, A. F. J. (2013). Comparison of the Hirsch-index with standard bibliometric indicators and with peer judgment for 147 chemistry research groups. Scientometrics, 67(3), 491–502. doi:10.1556/Scient.67.2006.3.10

Van Raan, A. F., Van Leeuwen, T. N., & Visser, M. S. (2011). Severe language effect in university rankings: Particularly Germany and France are wronged in citation-based rankings. Scientometrics, 88(2), 495-498.

Van Raan, A. F., Van Leeuwen, T. N., Visser, M. S., Van Eck, N. J., & Waltman, L. (2010). Rivals for the crown: Reply to Opthof and Leydesdorff. Journal of Informetrics, 4(3), 431-435.

Van Raan, A., Moed, H. F., & van Leeuwen, T. N. (2007). Scoping study on the use of bibliometric analysis to measure the quality of research in UK higher education institutions. Leiden, Netherlands. Retrieved from http://heer.qaa.ac.uk/SearchForSummaries/Summaries/Pages/RES75.aspx

Vanclay, J. K. (2012). Impact factor: Outdated artefact or stepping-stone to journal certification? Scientometrics, 92(2), 211-238.

Vaughan, L., & Hysen, K. (2002). Relationship between links to journal Web sites and Impact Factors. Aslib Proceedings: New Information Perspectives, 54(6), 356-361.

Vaughan, L., & Shaw, D. (2003). Bibliographic and web citations: What is the difference? Journal of the American Society for Information Science and Technology, 54(14), 1313-1322.

Vaughan, L., & Shaw, D. (2005). Web citation data for impact assessment: A comparison of four science disciplines. Journal of the American Society for Information Science and Technology, 56(10), 1075-1087.

Vaughan, L., & Shaw, D. (2008). A new look at evidence of scholarly citation in citation indexes and from web sources. Scientometrics, 74(2), 317-330.

Vaughan, L., & Thelwall, M. (2003). Scholarly use of the Web: What are the key inducers of links to journal Web sites? Journal of the American Society for Information Science and Technology, 54(1), 29-38.

Vieira, E. S., & Gomes, J. A. (2009). A comparison of Scopus and Web of Science for a typical university. Scientometrics, 81(2), 587-600.

Vieira, E. S., & Gomes, J. A. (2011). The journal relative impact: An indicator for journal assessment. Scientometrics, 89(2), 631-651.

Vinkler, P. (2007). Eminence of scientists in the light of the h-index and other scientometric indicators. Journal of Information Science, 33(4), 481-491.

Vinkler, P. (2010). The evaluation of research by scientometric indicators. Chandos Publishing.

Vinkler, P. (2012). The case of scientometricians with the “absolute relative” impact indicator. Journal of Informetrics, 6(2), 254-264.

Wager, E., Godlee, F., & Jefferson, T. (2002). How to Survive Peer Review. London: BMJ Books. doi:10.1258/jrsm.95.11.571-a

Wagner, C. S., & Leydesdorff, L. (2012). An Integrated Impact Indicator: A new definition of ‘impact’ with policy relevance. Research Evaluation, 21(3), 183-188.

Walters, W. H. (2007). Google Scholar coverage of a multidisciplinary field. Information Processing and Management, 43(4), 1121-1132.

Waltman, L. (2012). An empirical analysis of the use of alphabetical authorship in scientific publishing. Journal of Informetrics, 6(4), 700-711.

Waltman, L., & Costas, R. (2014). F1000 recommendations as a potential new data source for research evaluation: A comparison with citations. Journal of the Association for Information Science and Technology, 65(3), 433-445.

Waltman, L., & Schreiber, M. (2013). On the calculation of percentile-based bibliometric indicators. Journal of the American Society for Information Science and Technology, 64(2), 372-379.

Waltman, L., & Van Eck, N. J. (2012a). The inconsistency of the h-index. Journal of the American Society for Information Science and Technology, 63(2), 406-415.

Waltman, L., & Van Eck, N. J. (2012b). A new methodology for constructing a publication-level classification system of science. Journal of the American Society for Information Science and Technology, 63(12), 2378-2392.

Waltman, L., & Van Eck, N. J. (2013a). Source normalized indicators of citation impact: An overview of different approaches and an empirical comparison. Scientometrics, 96(3), 699-716.

Waltman, L., & Van Eck, N. J. (2013b). A systematic empirical comparison of different approaches for normalizing citation impact indicators. Journal of Informetrics, 7(4), 833-849.

Waltman, L., & Yan, E. (in press). PageRank-related methods for analyzing citation networks. In Y. Ding, R. Rousseau, & D. Wolfram (Eds.), Measuring scholarly impact. Springer.

Waltman, L., Calero-Medina, C., Kosten, J., Noyons, E., Tijssen, R. J., Eck, N. J., ... & Wouters, P. (2012a). The Leiden Ranking 2011/2012: Data collection, indicators, and interpretation. Journal of the American Society for Information Science and Technology, 63(12), 2419-2432.

Waltman, L., Van Eck, N. J., & Van Raan, A. F. (2012b). Universality of citation distributions revisited. Journal of the American Society for Information Science and Technology, 63(1), 72-77.

Waltman, L., Van Eck, N. J., Van Leeuwen, T. N., & Visser, M. S. (2013). Some modifications to the SNIP journal impact indicator. Journal of informetrics, 7(2), 272-285.

Waltman, L., Van Eck, N. J., Van Leeuwen, T. N., Visser, M. S., & Van Raan, A. F. (2011a). Towards a new crown indicator: Some theoretical considerations. Journal of Informetrics, 5(1), 37-47.

Waltman, L., Van Eck, N. J., Van Leeuwen, T. N., Visser, M. S., & Van Raan, A. F. (2011b). Towards a new crown indicator: An empirical analysis. Scientometrics, 87(3), 467-481.

Wan, J. K., Hua, P. H., & Rousseau, R. (2007). The pure h-index: Calculating an author's h-index by taking co-authors into account. COLLNET Journal of Scientometrics and Information Management, 1(2), 1-5.

Wan, J. K., Hua, P. H., Rousseau, R., & Sun, X. K. (2010). The journal download immediacy index (DII): Experiences using a Chinese full-text database. Scientometrics, 82(3), 555-566.

Wang, X., Wang, Z., & Xu, S. (2013). Tracing scientist's research trends realtimely. Scientometrics, 95(2), 717-729.

Wardle, D. A. (2010). Do 'Faculty of 1000 (F1000) ratings of ecological publications serve as reasonable predictors of their future impact? Ideas in Ecology and Evolution, 3, 11–15. doi:10.4033/iee.2010.3.3.c. http://library.queensu.ca/ojs/index.php/IEE/article/viewFile/2379/2479

Warner, J. (2000). Critical review of the application of citation studies to the Research Assessment Exercises. Journal of Information Science, 26(6), 453–460

Wei, C., Chen, Y., Yang, C., & Yang, C. C. (2010). Understanding what concerns consumers: A semantic approach to product feature extraction from consumer reviews. Information Systems and e-Business Management, 8(2), 149-167.

Weingart, P. (2005). Impact of bibliometrics upon the science system: Inadvertent consequences? Scientometrics, 62(1), 117-131.

Weiss, A., & James, R. (2013a). An examination of massive digital libraries' coverage of Spanish language materials: Issues of multi-lingual accessibility in a decentralized, mass-digitized world. Paper presented at the Proceedings – 2013 International Conference on Culture and Computing, Culture and Computing 2013, 10-14.

Weiss, A., & James, R. (2013b). Assessing the coverage of Hawaiian and pacific books in the Google Books digitization project. OCLC Systems and Services, 29(1), 13-21.

Weiss, D. (2005). Measuring success of open source projects using web search engines. In Scotto M., Giancarlo S. (Eds.): Proceedings of the 1st International Conference on Open Source Systems, Genova, Italy, pp.93-99. http://eprints.lincoln.ac.uk/76/1/oss2005-dweiss-projects-popularity.pdf

Weller, A. C. (2001). Editorial peer review: Its strengths and weaknesses. Medford, N.J: Information Today.

Weller, K., Dröge, E., & Puschmann, C. (2011). Citation analysis in Twitter. Approaches for defining and measuring information flows within tweets during scientific conferences. In M. Rowe et al. (Eds.), Proceedings of Making Sense of Microposts (#MSM2011). CEUR Workshop Proceedings. (pp.1–12), Heraklion, Greece.

Welpe, I. (2014). Incentives and performance : governance of research organizations.

Wen, X., Lin, Y., Trattner, C., & Parra, D. (2014). Twitter in academic conferences: Usage, networking and participation over time. Paper presented at the HT 2014 – Proceedings of the 25th ACM Conference on Hypertext and Social Media, 285-290.

Wennerås, C., & Wold, A. (1997). Nepotism and sexism in peer-review. Nature, 387, 341–343.

Wessely, S. (1998). Peer review of grant applications: what do we know? Lancet, 352(9124), 301–5. doi:10.1016/S0140-6736(97)11129-1

West, J. D., Bergstrom, T. C., & Bergstrom, C. T. (2010a). The eigenfactor metrics: A network approach to assessing scholarly journals. College and Research Libraries, 71(3), 236-244.

West, J., Bergstrom, T., & Bergstrom, C. T. (2010b). Big Macs and eigenfactor scores: Don’t let correlation coefficients fool you. Journal of the American Society for Information Science and Technology, 61(9), 1800-1807.

White, H.D., Boell, S.K., Yu, H., Davis, M., Wilson, C.S., & Cole, F.T. (2009). Libcitations: A measure for comparative assessment of book publications in the humanities and social sciences. Journal of the American Society for Information Science and Technology, 60(6), 1083-1096.

Whitley, R. (1984). The intellectual and social organization of the sciences. Oxford: Clarendon Press.

Whitley, R. (2000). The Intellectual and Social Organization of the Sciences. (Second Edition). Oxford: Oxford University Press.

Whitley, R. (2011). Changing Governance and Authority Relations in the Public Sciences. Minerva, 49(4), 359–385. doi:10.1007/s11024-011-9182-2

Whitley, R., & Gläser, J. (eds.) (2007). The Changing Governance of the Sciences. Dordrecht: Springer.

Whitley, R., Gläser, J., & Engwall, L. (eds.) (2010). Reconfiguring knowledge production : changing authority relationships in the sciences and their consequences for intellectual innovation. Oxford: Oxford University Press.

Whitley. R. (1984). The Intellectual and Social Organization of the Sciences. Oxford: Oxford University Press.

Wildgaard, L., Schneider, J. W., & Larsen, B. (2014). A review of the characteristics of 108 author-level bibliometric indicators. Scientometrics, 101(1), 125-158.

Wilkinson, D., Sud, P., & Thelwall, M. (2014). Substance without citation: Evaluating the online impact of grey literature. Scientometrics, 98(2), 797-806.

Willmott, H. (2011). Journal list fetishism and the perversion of scholarship: reactivity and the ABS list. Organization, 18(4), 429-442.

Wilson, J. D. (1978). Peer review and publication. Presidential address before the 70th annual meeting of the American Society for Clinical Investigation, San Francisco, California, 30 April 1978. The Journal of Clinical Investigation, 61(6), 1697–701. doi:10.1172/JCI109091

Wilson, M., & Holligan, C. (2013). Performativity, work-related emotions and collective research identities in UK university education departments: an exploratory study. Cambridge Journal of Education, 43(2), 223-241.

Woelert, P. (2013). The 'Economy of Memory': Publications, Citations, and the Paradox of Effective Research Governance. Minerva, 51(3), 341-362.

Wouters, P. (1999). The Citation Culture. University of Amsterdam, Amsterdam. Retrieved from http://garfield.library.upenn.edu/wouters/wouters.pdf

Wouters, P. (2014). The Citation: From Culture to Infrastructure. In: Cronin, B., & Sugimoto, C.R. (eds.), Beyond Bibliometrics: Harnessing Multidimensional Indicators of Scholarly Impact (pp. 47-66). Cambridge MA: MIT press.

Wouters, P., & Costas, R. (2012). Users, narcissism and control: Tracking the impact of scholarly publications in the 21st century. SURFfoundation. Retrieved November 29, 2014, from https://www.surf.nl/kennis-en-innovatie/kennisbank/2012/rapport-users-narcissism-and-control.html

Wouters, P., Bar-Ilan, J., Thelwall, M., Aguillo, I. F., Must, Ü., Havemann, F., … & Schneider, J. (2010). Academic Careers Understood through Measurement and Norms (ACUMEN) (pp. 1–39).

Wouters, P., Glänzel, W., Gläser, J., & Rafols, I. (2013). The dilemmas of performance indicators of individual researchers: an urgent debate in bibliometrics. ISSI Newsletter, 9(3), 48–53.

Wuchty, S., Jones, B. F., & Uzzi, B. (2007). The increasing dominance of teams in production of knowledge. Science, 316, 1036-1039.

Zahedi, Z., Costas, R., & Wouters, P. (2014). How well developed are altmetrics? A cross-disciplinary analysis of the presence of 'alternative metrics' in scientific publications. Scientometrics. 101(2), 1491-1513.

Zahedi, Z., Haustein, S. & Bowman, T. (2014). Exploring data quality and retrieval strategies for Mendeley reader counts. Presentation at SIGMET Metrics 2014 workshop, 5 November 2014. Available: http://www.slideshare.net/StefanieHaustein/sigme-tworkshop-asist2014

Zhang, L., & Glänzel, W. (2012). Proceeding papers in journals versus the "regular" journal publications. Journal of Informetrics, 6(1), 88-96.

Zimmermann, C. (2013). Academic rankings with RePEc. Econometrics, 1(3), 249-280.

Zitt, M. (2010). Citing-side normalization of journal impact: A robust variant of the audience factor. Journal of Informetrics, 4(3), 392-406.

Zitt, M., & Small, H. (2008). Modifying the journal impact factor by fractional citation weighting: The audience factor. Journal of the American Society for Information Science and Technology, 59(11), 1856-1860.

Zitt, M., Ramanana-Rahary, S., & Bassecoulard, E. (2003). Correcting glasses help fair comparisons in international science landscape: Country indicators as a function of ISI database delineation. Scientometrics, 56(2), 259-282.

Zitt, M., Ramanana-Rahary, S., & Bassecoulard, E. (2005). Relativity of citation performance and excellence measures: From cross-field to cross-scale effects of field-normalisation. Scientometrics, 63(2), 373-401.

Zuccala, A., & Guns, R. (2013). Comparing book citations in humanities journals to library holdings: Scholarly use versus "perceived cultural benefit" (RIP). In Paper Presented at the Proceedings of ISSI 2013—14th International Society of Scientometrics and Informetrics Conference (pp.353–360). Vienna, Austria: AIT Austrian Institute of Technology GmbH Vienna.

Zuccala, A., Guns, R., Cornacchia, R., & Bod, R. (in press, 2014). Can we rank scholarly book publishers? A bibliometric experiment with the field of history. Journal of the Association for Information Science and Technology. http://www.illc.uva.nl/evaluating-humanities/RankingPublishers(Preprint_2014).pdf

Zuckerman, H., & Robert Merton. (1971). Patterns of Evaluation in Science : Institutionalisation, Structure and Functions of the Referee System. Minerva, 9(1), 66–100.

Appendix A: Comparisons between GS and conventional citation indexes

Article	Dataset / discipline	Main results	Relevant conclusions for the REF
Bauer & Bakkalbasi (2005)	Articles from the Journal of the American Society for Information Science and Technology (JASIST) published in 1985 (41 papers) and 2000 (105 papers).	GS retrieved 4.5 and 3.9 times more citations than did WoS and Scopus, respectively, for papers published in 2000. However, WoS citations were 8.7 times higher than GS for older papers published in 1985.	'A search of Google Scholar will likely reveal both traditional journal articles, some of which will also be covered in Web of Science and Scopus, and additional unique material, but the scholarly value of some of the unique material remains an open question.' (No Page, online)
Meho & Yang (2007)	Over 1,457 publications by 25 library and information researchers	GS located 53% more citations than the union of WoS and Scopus and increased the total number of citations by 93%. There were significant correlations between GS and both WoS (0.874) and Scopus (0.970).	'GS stands out in its coverage of conference proceedings as well as international, non-English language journals.' (Page: 2105).
Kousha & Thelwall (2007a)	A sample of 1,650 journal articles published in 2001 in science and the social sciences (biology, chemistry, physics, computing, sociology,economics, psychology, and education)	GS citations were more numerous than WoS citations in economics (769%), education (507%) computer science (201%), sociology (219%), psychology (200%), but not in science excluding computer science (201%). GS citations highly correlated with WoS citations across all fields (from 0.825 in biology to 0.551 in education).	'There are clear disciplinary differences between conventional and Web-based citation patterns… and Google Scholar is a more comprehensive tool for citation tracking for social science. However, the quality of sources of citations (citing documents) retrieved by

			Google Scholar is an important factor to take into account' (Page: 1063-1064)
Mayr & Walter (2007)	9,500 journals from five databases, Thomson Scientific (SCI, SSCI, AH), Directory of Open Access Journals (DOAJ) and German social sciences literature (SOLIS)	About 86%, 88% and 80% of WoS-indexed journals in SCI, SSCI and AH were identified in GS searches (January 2007), respectively. About 68% of DOAJ journals and 70% of SOLIS journals were found in GS.	'The study shows that the majority of the journals on the five lists queried can be retrieved in Google Scholar…The international journals from the Thomson Scientific List (particularly from the area of STM) are fairly well covered.' (Page: 828). However, its coverage of the DOAJ list and German literature was lower than that of Thomson Scientific databases.
Meho & Rogers (2008)	22 top human-computer interaction researchers	Average GS h-indexes (20.6) were higher than for Scopus (12.3) and WoS (8.0) and there was a significant correlation (Spearman 0.960) between the GS h-index and h-indexes for Scopus and WoS.	'The main difference between the two rankings is that Google Scholar helps distinguish between the researchers in a more nuanced fashion than the union of Scopus and Web of Science, as evidenced by the larger variance between top-ranked and bottom-ranked researchers' (Page: 1724).
Vaughan & Shaw (2008)	A sample of 1,483 publications of library and information science	GS citation medians (ranging from 1-3) were significantly higher than WoS citation medians (zero for all types of publications except for	'In its current incarnation, Google Scholar has problems. Citing and cited papers are confused; and a

	faculty	books with median 1). Significant correlations between WoS and GS citations and manual (0.43 to 0.75 depending on the type of publication) and checking of citing GS citing sources revealed that about 92% of GS citations represent intellectual impact (e.g. formal citations).	single citation act may be represented multiple times when one citing work appears on several web pages. In spite of these problems, Google Scholar is a promising tool for research evaluation. If the current, beta, version of Google Scholar evolves in the right direction, it could be a serious challenger to WoS.' (Page: 328)
Bar-Ilan (2008)	47 highly cited Israeli researchers and three Nobel Prize winners	In many cases h-indexes of highly-cited Israeli researchers from GS were higher than from WoS and Scopus, especially for mathematicians and computer scientists. The average number of citations that the top h documents received in GS (153) was much higher than in WoS (21).	'The findings show that it matters which citation tool is used to compute the h-index of scientists. Also there seems to be disciplinary differences in the coverage of the databases. The differences in citation counts create a dilemma for science policy makers and promotion committees.' (Page: 269)
Kousha & Thelwall (2008)	A sample of 882 articles from 39 ISI-indexed journals in 2001 from biology, chemistry, physics and computing	43% of GS citations were also in WoS, although there were disciplinary differences. OA articles from non-WoS journals (34.5%), conference papers (25.2%), and e-prints/preprints (22.8%) were the most common sources of GS unique citations and the majority of GS unique citations (70%) were from full-text documents.	GS seems a useful tool 'for researchers using the citation tracking capability of Google Scholar for selecting a wider range of citations for their own work and non-evaluative purposes… However, the minimal amount of information known about Google Scholar's contents suggests caution for those

			seeking to use its citation data for research evaluation.' (Page: 290)
Bornmann et al. (2009)	1,837 articles accepted for publication in the journal Angewandte Chemie International Edition	Median citations for accepted papers derived from WoS SCI, Scopus and Chemical Abstracts (23, 23 and 25 respectively) were much higher than GS citation counts (one). This was due to poor coverage of the citing articles that GS couldn't access through the fee-based database providers at the time of the study.	They concluded that the results for in the field of chemistry 'on the one hand, the convergent validity of citation analyses based on data from the fee-based databases and, on the other hand, the lack of convergent validity of the citation analysis based on the GS data.' (Page: 33). However, GS citations might be beneficial for the fields of engineering, computer science and mathematics, social sciences, arts and humanities, where wider publication types are needed for citation analysis.
Kulkarni et al. (2009)	328 articles published in *JAMA*, *The Lancet*, or the *New England Journal of Medicine* (October1999-March 2000)	The GS citation median (160) was higher than Scopus (149) and WoS (122). GS retrieved a median of 37% more citations for *JAMA*, 32% for *The Lancet*, and 30% for *NEJM* articles than WoS.	'Web of Science, Scopus, and Google Scholar produced quantitatively and qualitatively different citation counts for articles published in 3 general medical journals. In offering alternative scopes of coverage and search algorithms, new citation databases raise questions of how to count citations.

			For example, should a citation on a non–peer-reviewed Web page be viewed as quantitatively equivalent to a citation in a high-profile peer-reviewed medical journal?' (Page: 1096)
Franceschet (2010a)	A sample of the publications of a group of Italian computer science scholars	GS extracted metrics were much higher than those from WoS: five times higher for paper-based indicators, eight times for citation-based indicators and three times for h type indicators. There were significant correlations between GS and WoS citation indicators (for citations 0.92 and for the h-index 0.65)	'…Great care must be taken when selecting the data source for the analysis. Our advice here is to perform a (time-consuming) join of the publications and citations contained in the two databases and use the combined universe to compute the h-index for scholars and journals.' (Page: 257)
Henzinger et al. (2009)	5,283 computer scientists and 1,354 physicists in WoS and GS	The average h-index derived from GS was 3.54 for computer scientists and 2.19 from WoS. In contrast, for physicists the average h-index in WoS (7.15) was slightly higher than for GS (6.70), although in both fields the GS citation medians were higher than those of WoS.	They concluded that 'wherever possible at least two different databases should be consulted and the relative ranking should only be trusted if it is consistent between the databases'. (Page: 473)
Lasda Bergman (2012)	The top five journals ranked highest by the 556 faculty	GS citations were more frequent than (3,272) Scopus (2,126) and WoS (1,741). About 44% of GS	'Google Scholar may not be as reliable as either Scopus or Web of Science

	members surveyed in the field of social work	citations were neither in WoS nor Scopus, whereas only 25% of citations of both WoS and Scopus were in GS. The overlap between GS and both WoS and Scopus was about 31%.	as a stand-alone source for citation data. Nonetheless, to obtain the most comprehensive citation count, one must use all three resources.' (Page: 378)
Mikki (2010)	Publications of 29 earth sciences authors	Just under 70% of the publications found were in GS alone, in contrast to 5% for publications in WoS alone. Nevertheless, citation and h-index values of common publications for authors were almost identical in the two databases. There was a high correlation between GS and WoS citation counts (0.74)	'The amount of earth science content is comprehensive in Google Scholar. It covers about 85% of content indexed by ISI WoS.' (Page: 330)
Mingers & Lipitakis (2010)	Over 4,600 publications in business and management from three UK business schools 2001-2007.	Just under half (48%) of the publications by the business and management researchers were in WoS, whereas GS searches found 66% of them (including about 90% of the journal articles). GS mean citations per paper were almost twice as much as WoS, although there were disciplinary differences.	They concluded that because WoS includes less than half of the journals, papers and citations found by GS, it is reasonable to use GS for impact assessment of scholars in business and management. 'Web of Science should not be used for measuring research impact in management'. (Page: 613)
Kousha et al. (2011)	1,000 books submitted to the 2008 UK Research Assessment Exercise (RAE) in seven book-based	GS citations to books were 3.2 times more common than Scopus citations and their medians were more than three times as high as Scopus median citations (medians of 13 and four respectively). There	In terms of practical implications for the UK, 'the absence of a plan to use citation information to inform expert reviewers about the impact of

	disciplines (archaeology, law, politics and international studies, philosophy, sociology, history, and communication, cultural and media studies)	were strong correlations between GS and Scopus citations to books (ranging from 0.744 in history to 0.833 in sociology). Based on a sample of 100 books, GS retrieved 84% unique citations that were not in Scopus, whereas the corresponding figure for Scopus was 45%.	research outputs in the REF in the arts, humanities and a number of other panels…, may be a drawback in quality assessment of UK research because of the difficulty in assessing large numbers of books."'(Page: 16)
Amara & Landry (2012)	Faculty members in Canadian business schools	For GS the mean number of publications (21.6), citations (271.5), and the h-index (4.6) were much higher than WoS (5, 50.8 and 1.9, respectively). High significant correlations between WoS and GS were found for contributions (0.793), citations (0.819), and h-indexes (0.815).	In the field of business and management, universities or other agencies 'should complement the data provided in WoS with those provided in GS.' (Page: 554)
De Groote & Raszewski (2012)	The publications of 30 College of Nursing faculty	H-indexes extracted from GS, WoS and Scopus strongly correlated with each other (0.835 for GS, WoS and 0.830 for GS and Scopus and 0.869 for WoS and Scopus). GS provided the highest h-indexes and more unique citations than Scopus and WoS (1312, 250 and 93 unique citations, respectively).	'More than one tool should be used to calculate the h-index for nursing faculty because one tool alone cannot be relied on to provide a thorough assessment of a researcher's impact. If nursing researchers are interested in the most comprehensive individual h-index, several databases should be searched to obtain the most comprehensive list of citing articles.'(Page: 391)
Minasny et al. (2013)	340 early-career and highly-cited soil researchers from all	The number of papers and citations to them in GS were 2.3 and 1.9 times higher than for WoS	'There is a large difference between the number of citations, number of

	over the world.	for soil researchers. High correlation between the h-index of the researchers using GS, WoS and Scopus.	publications and the h-index using the three databases.' (Online)
De Winter et al. (2014)	Two highly-cited classic articles and 56 articles from different subjects.	The retroactive growth of GS citation (median of 170%) was considerably higher than WoS (median of 2%). The actual growth of GS was also slightly higher than WoS (54% and 41%, respectively).	'GS has exhibited a striking retroactive expansion, considerably increasing its coverage of scientific literature as compared to 1 year after its inception. It is possible that GS fully covers WoS in the foreseeable future. However, improved metadata, more sophisticated search functions, and a stricter control against citation manipulation are challenges for GS yet to be met.' (Page: 1562)

Appendix B: Sources of data for alternative impact assessment

Alternative metric / Source	**Advantages**	**Limitations**
Most types Manual collection	Free. Data collection is transparent to the collector.	Time consuming. Needs expertise to identify all correct matches in some cases.
A range of altmetrics Commercial providers	Altmetric.com, ImpactStory.org and PlumAnalytics.org provide a range of altmetrics. May provide free access to data for individual researchers.	Data gathering methods may not be fully transparent.
GS citations Publish or Perish http://www.harzing.com/pop.htm	Automatically generates bibliometric indictors based on GS data including total citations and average citations per paper, author and year in addition to different h-indexes. Users can import external data from WoS, Scopus or several citation formats such as EndNote, RefMan or RIS format.	It may not be possible to use the software for bulk downloading citation data. The search results might differ from live GS search. Searching common authors' names (e.g. Taylor Smith) may retrieve publications by other authors and this needs extensive manual checking to exclude irrelevant authors and it does not limit results based on affiliations.
Twitter GNIP or DataSift	Deliver all Twitter matches to any given query.	May be expensive, especially for multiple queries and queries covering a long period of time.
Libcitations WorldCat API https://www.worldcat.org	Provides automatic data collection for bibliographic records and information about libraries that own the items and links to online catalogue records.	Not all world libraries are included in the database. Use of the WorldCat Search API is presently limited to employees of qualifying institutions that both contribute their library holdings to WorldCat and participate in the WorldCat.org programme.
GB citations	Performs automatic searches to locate	Although the software gives high overall

Webometric Analyst http://lexiurl.wlv.ac.uk	citations from online books indexed by GB. It filters out false matches and unwanted results from advertisements, book reviews, and bibliographies and so on. Data can be exported from WoS, Scopus or other standard sources such as WorldCat.org for query building. It is possible to add formal citation counts from WoS or Scopus query files for comparisons with GB citations. Free.	accuracy and coverage for the automatic GB citation searches (over 90%), the filtering is all based upon heuristics and so the results may be poor for books with very short or general book titles. May not be able to cope with huge numbers of books.
Online course syllabi for educational impact Webometric Analyst	Performs specific queries to locate citations in online syllabi or course reading lists via Bing. Limits syllabus mentions to the world university websites and excludes irrelevant results based on different rules. The data input could be from WoS, Scopus or other bibliographic sources for generating queries.	Many academic syllabi are not indexed by Bing and the method embedded in the software does not fully cover non-English syllabi and non-academic websites. The filtering technique is heuristic-based and may give false matches. Recall and precision of academic syllabus mentions are largely dependent on queries for capturing educational impact. Only the first 5,000 queries per month via Bing are free.
Citations from Academic Presentations Webometric Analyst	Citations from online presentations can be automatically located through specific queries limiting search results to presentation file types (e.g. ppt or pptx) via Bing. Bibliographic data can be used for searches from WoS, Scopus or other standard bibliographic sources.	Bing probably covers a smaller fraction of the web than does Google and there is no filtering method to exclude irrelevant presentations. Hence, using absolute mention counts could be problematic. A maximum of 5,000 queries per month via Bing are free.
Mendeley readership bookmarks Webometric Analyst	Uses the Mendeley API to collect readership counts. It generates queries based on bibliographic data from articles (article title, authors, year) and reports different statistics about Mendeley readership such as positions of registered readers.	Ignores data for duplicate Mendeley records for the same publication.
Amazon book	Uses Bing to locate books in Amazon (e.g. last name of the first author, the	Needs manual input to guide the process. Needs extra manual searching

reviews Webometric Analyst	book title and ISBN); downloads the Amazon book webpages and extracts the number of customer reviews, the average ratings of the reviews and the Amazon best sellers rank from each book.	to identify missing results.
Book review sentiments SentiStrength	Estimates and reports the strength of positive and negative sentiment in short texts.	Reports sentiments about the content of the books reviews rather than their overall judgments.
YouTube science videos Webometric Analyst	Collects data and reports statistics from YouTube videos including the number of views, comments, likes, dislikes via the YouTube API.	

List of Abbreviations and Glossary

AERES	Agency for Evaluation Research and Higher Education
ARC	Australian Research Council
ARL	Association of Research Libraries
BKCI	Book Citation Index
DOI	Digital Object Identifier
DUI	Data Usage Index
EPO	European Patent Office
ERC	European Research Council
ESF	European Science Foundation
ESRC	Economical and Social Research Council
GB	Google Books
GP	Google Patents
GS	Google Scholar
GSC	Google Scholar Citations
HEFCE	Higher Education Funding Council for England
JIF	Journal Impact Factor

Matthew Effect In sociology, the Matthew effect (or accumulated advantage) is the phenomenon where "the rich get richer and the poor get poorer. It derives its name from the biblical Gospel of Matthew (Matthew 25:29, King James Version), http://en.wikipedia.org/wiki/Matthew_effect

OCLC	Online Computer Library Center
PLoS	Public Library of Science
RAE	Research Assessment Exercise
REF	Research Excellence Framework
SJR	SCImago Journal Rank
SNIP	Source Normalised Impact per Paper
SSH	Social Sciences and Humanities
STEM	Science, Technology, Engineering and Mathematics
URL	Uniform Resource Locator
USTPO	United States Patent and Trademark Office
WoS	Web of Science